Mind-space and individual differences in theory of mind

Conway, Jane

Awarding institution:
King's College London

Mind-space and Individual Differences in Theory of Mind

Jane Rebecca Mary Conway

Social, Genetic and Developmental Psychiatry Centre, Institute of Psychiatry, Psychology and Neuroscience, King's College London

Submitted for the degree of Doctor of Philosophy 2019

Abstract

Understanding the social world requires making accurate inferences about the contents of other people's minds, being able to represent in one's own mind the thoughts, beliefs, and intentions of another. With a long history of investigation in Philosophy, the ability to represent others' mental states has been the subject of considerable scientific effort in Psychology and Cognitive Neuroscience for 40 years (where it is known as 'theory of mind'). However, this large body of work has produced few ideas of how to conceptualise individual differences in theory of mind ability. This thesis presents a theoretical framework for studying such variation.

In this work, a distinction is made between minds and mental states whereby the term 'mind' refers to an individual's complete set of cognitive systems, and the term 'mental state' refers to the representational content generated by that set of systems. As mental states are a product of the minds that generate them, accurate inference of another's mental states is likely aided by representing multiple features of minds and variability between minds. Chapter 2 introduces the 'Mind-space' theory, which suggests that minds are represented in a multidimensional space and the probability of specific mental states is dependent on location in this space. In the Mind-space framework, individual differences in the representation of other minds, and in the accuracy of mental state inferences, are attributable to the properties of the space, the ability to locate a target mind in the space, and the mappings between location in space and mental state probabilities. Chapter 3 presents four experiments that provide empirical support for these predictions. Chapter 4 examines whether trait dimensions in Mind-space adapt following brief experience of populations with different trait distributions. Chapter 5 investigates the representation of another's memory performance – a nonsocial Mind-space dimension – and how this may be affected by the self's

performance and metacognitive accuracy of such. Chapter 6 first asks whether there exist domain-specific cognitive mechanisms for implicit mental state representation, and further assesses individual differences therein.

Chapter 2 discusses how the Mind-space theoretical framework addresses limitations in the previous literature to studying variation in theory of mind, and its relevance for understanding human development, intergroup relations, and the socio-cognitive impairments seen in several psychiatric and neurodevelopmental conditions. Chapter 7 summarises the studies presented in the thesis, and considers limitations and future applications of Mind-space for the study of individual differences in mind and mental state representation.

Acknowledgements

I am grateful for the support of the Economic and Social Research Council in granting me a studentship. I consider it a huge privilege to have been given the time to learn and to think. Thank you to the members of the Social, Genetic, & Developmental Psychiatry Centre at the IoPPN, Brasenose College Oxford and the Dept. of Experimental Psychology, it has been magical to spend time with people with astonishing amounts of knowledge and curiosity. I am thankful for the consistent help of Wijnand van Tilburg, Caitlin Patrick, and Barbara Maughan with the many practical aspects of conducting this research.

Our lab and its hive mind is a truly special community. I am hugely grateful to this group of individuals for their steadfast support, razor sharp thinking tools, and boundless sense of fun. Your contributions and challenges have shaped and improved this work. Thanks to Michel-Pierre Coll, Hannah Hobson, Rebecca Brewer, Annabel Nijof, Jenny Murphy, Tegan Penton, Mirta Stantic, David Plans, Eri Ichijo, and Hélio Clemente Cuve.

I am greatly indebted to my friends who have been there for me throughout. Special thanks to Mike and Katie Johnson and Cissie Fu, and to my fellow PhD adventurers Sophie Sowden, Hannah Pickard, Luke Norman, and Esra Yarar. Laurence Berry and Baocong Xia, you are missed. To Tinta Deasy, your enduring friendship is a constant source of joy and solace to me. To Kirsten Purves, for giving the best hugs. To Toby Wise, for being a source of wisdom on everything, including bowling. And to Laura Riddleston, for sharing an enthusiasm for bees and creating so much laughter.

I owe a huge amount to Caroline Catmur, who somehow manages to combine in one human being the ability to think in the most pure logic, an incredible eye for detail, and a huge heart that cares for her student's wellbeing. Thank you, Caroline, for your sage advice and all you have given to this work.

I have benefitted enormously from the kind advice and encouragement of Franky Happé. Thank you, Franky, it has been a delight to discuss beliefs about things with you, and I remain in awe of how elegantly you approach scientific theories and the leadership of an academic community.

I am immensely grateful to Geoff Bird. Geoff, how to express adequately (note the infinitive was not split there) my thanks to you? I am so incredibly lucky to have had the chance to work with you, to benefit from your scientific bravery in tackling literatures anew, and to enjoy the process. I greatly appreciated your guidance and advice, and especially the tremendous generosity you have shown in the giving of your time and cognitive resources. Thank you.

I wish to acknowledge the important role my family play in supporting me always. Thanks to Owen, Alan, Sarah, F., and my wonderful parents, Anne and Walter. Finally, to Ruairidh Howells, I will be forever grateful for the support you have given me throughout this thesis and beyond.

About this thesis

This thesis is presented as a thesis incorporating publications. Chapters 2 and 6 are exact copies of published journal articles, and are available on the relevant journal websites. Chapter 3 is presented in its form as a manuscript currently under review and has incorporated the invited revisions of the editor. Chapters 4 and 5, along with the general introduction and discussion chapters, were written specifically for this thesis. For consistency, each chapter has its own reference list. There are three videos related to this thesis, and these are available via the links provided.

Contents

Figures

Tables

1. Introduction

In 1978 Premack and Woodruff asked whether Sarah, a 14-year-old chimpanzee, had a 'theory of mind', that is, an ability to ascribe mental states to herself and others (Premack & Woodruff, 1978). Such mental states go beyond explanations of superficial behaviour by referring to what is inside the head, for instance thoughts, beliefs, intentions, and desires. In humans, these invisible mental states play an important societal role. The assumption that one can represent accurately in one's own mind the mental states of another person's mind underpins social relationships and institutions. For instance, in the justice system, members of a jury must not only consider the potentially illegal behaviour but the beliefs and intentions motivating it (Alicke, Mandel, Hilton, Gerstenberg, & Lagnado, 2015); in education, a teacher must represent what a child knows or does not know in order to be able to instruct them (Heyes & Frith, 2014; Csibra & Gergely, 2006); and in health, a symptom of many neurodevelopmental and psychiatric disorders is either difficulty in understanding others or a tendency to make unusual social inferences (Happé, 1994; Happé & Frith, 1996). Thus, Premack and Woodruff introduced an influential scientific concept that allows us to examine understanding of not just the external, observable, aspect of others' actions, but the internal cognitions driving those actions. A vast literature has since emerged, spanning developmental, comparative, evolutionary and clinical psychology, and cognitive neuroscience, robotics, and artificial intelligence (Wellman, Cross, & Watson, 2001; Krupenye, Kano, Hirata, Call, & Tomasello, 2016; Russell, Schmidt, Doherty, Young, & Tchanturia, 2009; Spunt & Adolphs, 2015; Asada, MacDonald, Ishiguro, & Kuniyoshi, 2001; Lake, Ullman, Tenenbaum, & Gershman, 2017; Rabinowitz et al., 2018).

1.1 Theories of Theory of Mind

1.1.1 Explicit vs. Implicit Theory of Mind

It has been proposed that there may be two cognitive systems for representing others' mental states - an explicit system and an implicit system (Frith & Frith, 2008). Explicit mental state representation is a slow, cognitively demanding, verbally-reportable process, which relies on other cognitive processes like language and memory. There is broad consensus that the explicit system is present in humans from the approximate age of 5 years (Wellman, Cross, & Watson, 2001). Such consensus is possible due to the use of verbal response measures that provide an unambiguous statement of what the inferred mental state is.

In contrast, there is much debate as to whether there exists a second system by which human adults, and even infants and nonhuman animals, can represent mental states – or simpler 'belief-like' states – implicitly (Apperly & Butterfill, 2009; Butterfill & Apperly, 2013). This implicit system is suggested to involve fast, subconscious processes that do not rely on general cognitive demands (Qureshi, Apperly, & Samson, 2010). The debate rests on the interpretation of experimental effects used to support claims of implicit mental state representation (Heyes, 2014a; Heyes, 2014b). These effects depend on ambiguous nonverbal measures such as reaction time or anticipatory looking (Senju, Southgate, Snape, Leonar, & Csibra, 2011), and this ambiguity has resulted in questions as to the reliability and validity of these measures as reflecting implicit mental state representation (Kulke, Johannsen, & Rakoczy, 2019; Dörrenberg, Rakoczy, & Liszkowski, 2018; Kulke, von Duhn, Schneider, & Rakoczy, 2018; Kulke & Rakoczy, 2018; Powell, Hobbs, Bardis, Carey, & Saxe, 2018; Kulke, Reiß, Krist, & Rakoczy, 2017). Some interpret these effects richly, as evidence of implicit mental state representation, while others suggest that low-level domain-general cognitive processes

underpin these effects (Heyes, 2014a; Heyes, 2014b). To adjudicate between mentalistic and non-mentalistic interpretations of implicit measures, it is necessary to empirically test these opposing hypotheses through experiments designed to juxtapose them (Heyes, 2015). Evidence for the existence of an implicit theory of mind would be significant due to how claims of such a system in human infants, chimpanzees, and birds, have been used to support the idea of an innate, evolutionarily-specified module (Leslie, 1992; Wang & Leslie, 2016).

1.1.2 Simulation Theory vs. Theory-Theory

In attempting to explain *how* others' mental states are inferred, two types of cognitive processing have been proposed: Simulation Theory (ST) and Theory-Theory (TT) (Carruthers & Smith, 1998). In ST, a person uses their own mind to run a simulation of the scenario another person is in to generate mental states the self would have, and then these mental states are ascribed to the other. In TT, a person has a body of knowledge that connects mental state concepts and the principles underlying their manifestations and interactions, analogous to a scientific theory (Gopnik & Wellman, 1992), which one applies to derive inferences about others' mental states. While some argue in favour of one theory over the other (Saxe, 2005), current consensus considers a hybrid account whereby both strategies are employed (Mitchell, 2005). Notably, ST supports explanations of egocentric bias and why the self may be useful in the absence of information about the other (Mitchell, 2009), whereas TT accounts for systematic biases that distinguish the self from other, or systematic errors in reasoning (Saxe, 2005). However, a significant challenge to both ST and TT lies in how they have performed as scientific theories in the theory of mind literature; they have not provided clear testable predictions to distinguish between them or to explicate fully mental state processing (Apperly, 2009).

1.1.3 Modelling Another Mind

Godfrey-Smith (2005, pp. 4) suggested "*one aspect of ordinary folk-psychological skill might best be described not as grasp of a theory but as something like facility with a model… we are bringing something like a model to bear on the person we are trying to interpret*". He draws on concepts in philosophy of science to use the term 'model' in a specific sense of theorising: briefly, model-building is one type of theorising, its goal is to create hypothetical structures to facilitate understanding of a target system, such understanding is attributable to the resemblance relations between the model and real system (pp. 2-3). Importantly, there can be different *construals* of a model, i.e. the extent to which it actually represents the target system; it may serve only as useful predictive tool without having any explanatory value for the true system, or it may be posited as a true account of how the system functions.

Godfrey-Smith distinguishes model-based understanding from theorising in the Theory-Theory account because theories in the classic and latter sense require laws, generalizations, and truth conditions whereas models are concerned with useful resemblance between the hypothetical and target system. Gopnik and Wellman (1992) seem to use a classical definition of 'theory', saying that the constituent entities are 'lawfully' interrelated with one another, and comparing children's theory of mind to physical theories of gravity and planetary motion. However, such laws are elusive in explicit everyday ascriptions of mental states (Godfrey-Smith, 2005; Maibom, 2003). Godfrey-Smith further distinguishes model-based understanding from Simulation Theory by emphasising that *hypothetical* models are used in the former, whereas in the latter one's own mind is used as a *physical* model of another. Nevertheless, he does acknowledge that modelling may be a modification of TT, and, in the case of insufficient information, simulations may generate input to the model.

Godfrey-Smith's modelling theory of other minds has distinct promise to advance understanding of individual differences in mental state representation. Notably, he acknowledges that (1) mental state inferences lack definite truth conditions, and (2) the *model psychological profile* may vary depending on the target mind. This shifts the current emphasis on reporting the 'correct' mental state (Section 2.4) to considering the properties of the model a particular individual employs for a particular target mind, and the variation in models between individuals and indeed groups of individuals.

1.2 Measurement of Theory of Mind

When first encountering measures in the theory of mind literature, there are four interrelated types of heterogeneity that can mislead and confuse. First, the same cognitive process, usually defined as the ability to represent mental states, can be called by different names, including: mentalizing, mind reading, folk psychology, cognitive empathy, perspective-taking, in addition to theory of mind. Second, what constitutes a mental state varies considerably; this can range from visual perspectives (Samson, Apperly, Braithwaite, Andrews, & Bodely-Scott, 2010) to emotions (Baron-Cohen, Wheelwright, Hill, Raste, & Plumb, 2001) to propositional attitudes (Leslie, 1987). A third source is the same name being given to different cognitive processes. For instance, emotion recognition has been shown to be distinct from mental state ascription but both processes are commonly conflated in the literature (Oakley, Brewer, Bird, & Catmur, 2016). A fourth source is in how these various constructs are operationalized and measured. Methods range from the nonverbal (Senju et al., 2011) to vignettes (Happé, 1994) to neuroimaging (Richardson, Lisandrelli, Riobueno-Naylor, & Saxe, 2018). What is most concerning is that measures of theory of mind show poor convergence with one another (Warnell & Redcay, 2019); although the ability is invariably described as mental state representation, this corresponds to myriad processes which do not coalesce around a single construct as has been developed in other cognitive domains

such as Intelligence (Happé, Cook, & Bird, 2017). Neuroimaging work has proposed a 'theory of mind network' comprising regions that may serve different functions in theory of mind processing (Schaafsma, Pfaff, Spunt, & Adolphs, 2015). A meta-analysis has shown that different tasks activate different brain regions, but the bilateral temporoparietal junctions and medial prefrontal cortex are commonly activated (Schurz, Radua, Aichhorn, Richlan, & Perner, 2014). However, brain activation does not correspond with performance on theory of mind behavioural measures (Richardson et al., 2018; Dufour et al., 2013), highlighting an enduring problem in the literature of not being certain of exactly *what* is being measured.

1.3 Development of Theory of Mind

The vast majority of theory of mind research has focused on young children. As mentioned in Section 1.1.1 earlier, the use of nonverbal measures in human infants has supported claims of an ability to ascribe false beliefs to others from the age of 7 months (Kovács, Téglás, & Endress, 2010). These findings have been questioned in two ways: first, by recent studies that have not replicated (Kulke & Rakoczy, 2018); and second by explanations that attribute these effects to domain-general processes, such as retroactive interference or perceptual novelty (Philips et al., 2015; Heyes, 2014b). A further problem is in explaining why infants might be capable of ascribing false beliefs in nonverbal tasks, but fail to do so in verbal tasks at a much later age of 3 or 4 years (Ruffman, Garnham, Import, & Connolly, 2001). This contrast has led to a distinction between implicit and explicit cognitive systems, where there are signature limits on what can be processed by, and the flexibility of, the implicit system (Butterfill & Apperly, 2013; Apperly & Butterfill, 2009). For instance, it can only represent 'belief-like' states rather than metarepresentation of full propositional attitudes (Butterfill & Apperly, 2013; Apperly & Butterfill, 2009). This distinction between systems, along with the limits of the abilities of infants and children (or indeed adults), suggests that

the implicit system is not the acorn from which the explicit oak grows (Heyes & Frith, 2014).

Research on the typical development of explicit theory of mind has focused on the age at which children acquire different mental state concepts, with easier concepts mastered prior to more complex ones, for instance desires before beliefs (Wellman & Liu, 2004). A much-used measure is the Change-of-Location False Belief Task (Baron-Cohen, Leslie, & Frith, 1985), in which an agent leaves an object in a particular location before departing the scene and, while the agent is absent, the object moves to a new location; the critical test question asks where will this agent look for the object on their return? The 'correct' response is to report that the agent will look (and therefore believes the object to be) where they left it. This false belief is discordant with the true belief of the participant, who has observed the object moving, and the present state of the object at the time of testing. Children typically pass this test around the age of five years (Wellman, Cross, & Watson, 2001), but there is much debate as to why they fail it. Children tend not to make random errors, but rather ascribe their own egocentric true belief to the agent (Wellman, 2014). However, the task has concurrent general cognitive demands of language and executive functioning, indeed when these are reduced children can pass this test earlier (Rubio-Fernández & Geurts, 2013). Neuroimaging work suggests that passing the false belief test is associated with the maturation of white matter in, and functional connectivity of, the theory of mind brain network (Grosse Wiesmann, Schreiber, Singer, Steinbeis, & Friederici, 2017).

Throughout middle childhood, children's mental state reasoning becomes faster (Apperly, Warren, Andrews, Grant, & Todd, 2011), and more sophisticated by appreciating that people may have different beliefs even in the same situation (Lalonde & Chandler, 2002), and by incorporating more sources of information into their

mentalistic interpretations (further discussed in Chapter 3). Beyond childhood, changes in theory of mind are observed. Adolescents have been shown to make fewer egocentric errors than children but more than adults (Dumontheil, Apperly, & Blakemore, 2010). Into older adulthood (> 70 years old), studies have shown mixed results but an overall view indicates that theory of mind performance worsens in old age and this seems to correspond to decline in executive functions and in fluid (but not crystallized) intelligence (Moran, 2013).

The relative paucity of study of theory of mind in adults highlights a challenge in the field: the vast majority of research has focused on what it means to perform poorly on theory of mind measures, whether that be due to not having acquired a certain concept or to lacking the general cognitive functions necessary to support mental state reasoning. These two explanatory foci have limited the field to studying young children, and, to a limited extent, older adults who tend to have a degree of decline in general cognitive function. What is missing is the study of why people might become theory of mind experts, or 'better' at mentalistic interpretations, beyond these two explanations. Moreover, if such expertise depends on one's social and cultural experiences then it is likely that study of a generic theory of mind ability becomes less useful, rather the focus moves to a theory of the minds with which one has experience.

1.4 Group Differences in Theory of Mind

1.4.1 Autism Spectrum Disorder

Autism Spectrum Disorder (henceforth 'autism') is a genetically inherited neurodevelopmental disorder (Colvert et al., 2015). Those with autism show specific difficulties on theory of mind tasks (White, Hill, Happé, & Frith, 2009), and such difficulties have come to characterise what is in fact a heterogeneous condition (Happé, Ronald, & Plomin, 2006). The presence of this 'specific deficit' has been used in

arguments for the innate modularity of a theory of mind mechanism (Leslie, 1992; Wang & Leslie, 2016). However, the presence of a disorder with a specific theory of mind impairment is not sufficient for such claims as there are also genetically inherited neurodevelopmental disorders that affect culturally inherited skills, for example dyslexia and reading (Heyes & Frith, 2014).

Despite the many strengths of the theory of mind in autism literature, the presence of such a comparison group has inadvertently constrained the question to '*who* can pass theory of mind tasks?' rather than '*how* can theory of mind mechanisms be explained?'. For instance, people with autism can pass explicit theory of mind tasks, but at a more advanced verbal mental age than the age at which neurotypical children do so (Happé, 1995). This has led to a hypothesis that autistics use different or compensatory mechanisms compared to neurotypicals to logically 'hack out' the correct response (Livingston & Happé, 2017). However, without a full understanding of the mechanisms involved in neurotypicals, it remains only a hypothesis that autistics arrive at a mental state inference via a different cognitive route. The limited ceilings of current theory of mind measures mean that it is sometimes difficult to find differences from typical controls, especially for those with autism but without language or intellectual impairment (Murray et al., 2017). When these differences are observed, it is not clear how to interpret them.

For instance, in an advanced theory of mind video task, compared to matched neurotypical controls, autistic participants scored significantly lower on the accuracy of (1) what mental state they inferred and (2) their suggested social response – what they would say next in the interaction – but they used equivalent amounts of mental state words and metacognitive statements (Murray et al., 2017). This implies that they engaged in mentalistic interpretations but failed to report the response deemed correct.

Why their responses differed and how they were derived is not answered or answerable with current measures. The observed difference in scores was small: on average the autistic group scored 15.5/24 and the control group 18.8/24 (Murray et al., 2017). Although this was a statistically significant difference, it does not explain *why* their performance was worse on average. Furthermore, the lack of common error variance among people with autism (or indeed controls) highlights that focusing on summing the prescribed right or wrong responses does little to address how the mental state inferences were arrived at.

1.4.2 Cross-cultural Differences in Theory of Mind

While the search for evidence of theory of mind in chimpanzees and corvids has focused attention and efforts on its genetic evolution, less attention has been given to the rich role cultural evolution may play in shaping human theories of what is inside the heads of our conspecifics (Heyes & Frith, 2014). The criticisms of theory of mind measures thus outlined mean that tasks do not lend themselves well to the study of variation across individuals. Without a systemic understanding of variance, explanations of individual differences are limited to the most plausible *post hoc* account. For example, differences in the order of acquisition of mental state concepts are observed between cultures and this has been attributed to collectivist vs. individualist community practices (Shahaeian, Peterson, Slaughter, & Wellman, 2011; Slaughter & Perez-Zapata, 2014). When children in Samoa failed to pass reliably the false belief task by aged 10-12 years (Mayer & Träuble, 2013), this was posited to be due to an expectation that someone may move one's toys, which is a possibility when one lives in large extended family groups and open houses, as the Samoan children did.

The prescribed 'right' and 'wrong' task response may vary according to culture, but the role of cultural learning in, and cultural origins of, theory of mind have rarely

been addressed (Heyes & Frith, 2014). Moreover, a primary question is whether cultures actually have or use a theory of mind. Ethnography work has documented accounts of some cultures using few mental state terms and, considering the mind opaque, interpreting observable actions rather than internal thoughts (Lillard, 1998). Conceptions of minds and how they relate to agent's thoughts and intentions also vary cross-culturally, for instance different mind dimensions have been observed in North American vs. Fijian groups (Willard & McNamara, 2019). An insensitivity to cross-cultural variation, combined with the emphasis on theory of mind as a psychological universal, has resulted in expansive claims of "*the* structure of mind perception" (Gray, Gray, & Wegner, 2007) or of "discovering which dimensions *the brain* spontaneously uses to organize the domain of mental states" (Tamir, Thornton, Contreras, & Mitchell, 2016) in samples of, for example in the latter case, twenty people between the age of 18 and 27 years old in the Harvard University Study Pool (Tamir et al., 2016). The theory of mind literature needs to evolve in order to be capable of elucidating why and how cultures differ in their theories of minds. The majority of theory of mind measures present decontextualized minds in sparse scenarios, and therefore the literature suffers from the 'frame problem' (Dennett, 1984) of how to navigate to the relevant set within near limitless possible inferences; how membership of a 'community of minds' (Nelson, 2005) frames one's theories of minds is regrettably understudied.

1.5 Individual Differences in Theory of Mind

1.5.1 Correlations

A large focus of individual differences has been on the relationship between theory of mind ability and language (Milligan, Astington, & Dack, 2014) or executive functions including memory and inhibitory control (Lecce, Bianco, Devine, & Hughes, 2017; Devine, White, Ensor, & Hughes, 2016). Hughes and Devine (2015) have

proposed two accounts of individual differences: developmental lag, and genuine differences. Under the *developmental lag* account, differences are due to the rate of acquisition of different mental state concepts, but once these concepts are acquired, usually in childhood, this source of variance ceases. Under the *genuine differences* account, variation in theory of mind is attributed to correlates that endure beyond childhood and result in actual differences in ability, and they suggest that the evidence better supports this account.

Quantitative genetic studies can indicate the relative contribution of genetic and environmental factors influencing individual differences in theory of mind. Four studies have examined individual differences in theory of mind using a twin design (Hughes & Cutting, 1999; Hughes, Jaffee, Happé, Taylor, Caspi, & Moffitt, 2005; Ronald, Happé, Hughes, & Plomin, 2005; Ronald, Viding, Happé, & Plomin, 2006); the genetic heritability estimates ranged from 7-67%, the influence of the shared environment between 0-48%, and the nonshared environment between 32-66% (note that this estimate also includes measurement error). That the environment plays a role in theory of mind performance accords with associations between children's false belief understanding and family factors, including: socioeconomic status, parental mental state talk and mind-mindedness, and number of siblings (Hughes et al., 2005).

A bivariate analysis allows the examination of the genetic, shared and non-shared environmental contributions to the covariance between two traits (Plomin, DeFries, Knopik, & Neiderhiser, 2013). Due to the phenotypic correlation between theory of mind and language ability, all of the four studies described above included some measure of verbal ability and three of the studies conducted bivariate analyses. The first study concluded that 66% of the genetic influence on theory of mind was independent of genetic influence on verbal IQ (Hughes & Cutting, 1999). However, this

contrasts with the findings of the other two studies, perhaps due to the relatively smaller sample size, the younger age of the sample, or because of the measure of verbal ability used. In the study by Hughes et al. (2005) the genetic influences on theory of mind overlapped entirely with the genetic influences on verbal ability, as did the shared environmental influences on verbal ability with the shared environmental influences on theory of mind (Hughes et al., 2005, p. 361). The significant phenotypic correlation of .40 between theory of mind and verbal ability could be explained by factors common to both abilities, predominantly by genetic (41%) and shared environmental factors (56%) (Hughes et al., 2005). Similarly in the Ronald et al. (2005) study, verbal ability significantly predicted theory of mind ability, and the genetic correlation (.86) between the two abilities indicated that the genetic influences on both were almost entirely shared (Ronald et al., 2005, p. 421).

1.5.2 Ability vs. Propensity

The ability to make a mental state inference is often confounded with the propensity to do so. For instance, in the advanced theory of mind task discussed in Section 1.4.1, participants received lower scores for using less complex mental state language (Murray et al., 2017). This may reflect an inability to make complex inferences, or a lack of propensity to do so. As discussed in the Cross-cultural Section 1.4.2, the propensity to reference internal mental states can vary. The distinction between the fast and flexible implicit system and the slow, effortful explicit system becomes particularly relevant here, as the effort expended in explicit theory of mind processing will depend on an individual's motivation and tendencies. It has been demonstrated that, when motivated, both typical controls and autistic participants donate significantly more to charity when observed by another person, suggesting that they consider what that person thinks of them (Cage, Pellicano, Shah, & Bird, 2013). However, it may be that ability and propensity interact with one another. If one can

reference mental states, and if one is more prone to mentalistic interpretations of behaviour, this may result in greater accuracy over time.

1.6 Summary of the Limitations of the Existing Literature

1. To empirically test whether implicit theory of mind effects reflect domain-specific mental state processes or domain-general non-social processes, it is necessary to design experiments capable of distinguishing between these alternatives and, specifically, of providing positive evidence for one account rather than only negative evidence against the other.
2. Simulation Theory and Theory-Theory accounts have not provided comprehensive explanations of the mechanisms underpinning mental state processing.
3. There is an unhelpful degree of heterogeneity in the terms and measures used, most notably when the same name is given to different cognitive processes. This problem is further compounded by the lack of convergence and correlations between different measures, and the absence of any factor structure for the domains of social cognition.
4. Neuroimaging studies focusing on the theory of mind network show that brain activation does not reliably reflect behavioural task performance. As tasks tend to measure the 'correct' mental state inference, it is possible that participants are engaging in mental state representation but failing to report the prescribed response. Brain activation also does not provide a psychological account of how mental states are processed.
5. Existing theory of mind measures focus on determining the ages at which neurotypical groups pass a certain milestone or how clinical groups differ, but they perform poorly at accounting for individual differences. Moreover, there is

little conception of what it means to be 'better' at theory of mind beyond having acquired advanced concepts and having superior general cognitive abilities.

6. Individual differences to date have focused on correlates of theory of mind ability or the tendency to use it; both of these fall short of explaining how and why some individuals make different mental state inferences to other individuals. Furthermore, there exists no theoretical framework for individual differences in theory of mind that can generate testable predictions to account for such variation.

1.7 Thesis Aims and Outline

This thesis aims to address the problem of conceptualising individual differences in theory of mind. Chapter 2 presents 'Mind-space', a new theoretical framework for explaining why individuals may ascribe different mental states to different people, and why individuals may vary from one another in their ascriptions. It also presents testable predictions made by the theory. Mind-space begins to address the theoretical limitations described in the previous section 1.6. Chapters 3, 4 and 5 present six experiments directly testing predictions made by the Mind-space theory as described in Chapter 2. Finally, Chapter 6 presents three experiments that aim to test the hypothesis that implicit theory of mind involves domain-specific mentalistic processing (Limitation Number 1 in Section 1.6), and further examines individual differences therein.

1.8 References

Alicke, M. D., Mandel, D. R., Hilton, D. J., Gerstenberg, T., & Lagnado, D. A. (2015). Causal Conceptions in Social Explanation and Moral Evaluation: A Historical Tour. *Perspectives on Psychological Science*, *10*(6), 790–812. https://doi.org/10.1177/1745691615601888

Apperly, I. A., Warren, F., Andrews, B. J., Grant, J., & Todd, S. (2011). Developmental

continuity in theory of mind: Speed and accuracy of belief-desire reasoning in children and adults. *Child Development*, *82*(5), 1691–1703. https://doi.org/10.1111/j.1467-8624.2011.01635.x

Apperly, I. A. (2009). Alternative routes to perspective-taking: imagination and rule-use may be better than simulation and theorising. *The British Journal of Developmental Psychology*, *27*, 545–553. https://doi.org/10.1348/026151008X400841

Apperly, I. A, & Butterfill, S. A. (2009). Do humans have two systems to track beliefs and belief-like states? *Psychological Review*, *116*(4), 953–970. https://doi.org/10.1037/a0016923

Asada, M., MacDorman, K. F., Ishiguro, H., & Kuniyoshi, Y. (2001). Cognitive developmental robotics as a new paradigm for the design of humanoid robots. *Robotics and Autonomous Systems*, *37*(2–3), 185–193. https://doi.org/10.1016/S0921-8890(01)00157-9

Baron-Cohen, S., Leslie, A. M., & Frith, U. (1985). Does the autistic child have a "theory of mind"? *Cognition*, *21*(1), 37–46. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/9775957

Baron-Cohen, S., Wheelwright, S., Hill, J., Raste, Y., & Plumb, I. (2001). The "Reading the Mind in the Eyes" Test revised version: a study with normal adults, and adults with Asperger syndrome or high-functioning autism. *Journal of Child Psychology and Psychiatry, and Allied Disciplines*, *42*(2), 241–251. https://doi.org/10.1111/1469-7610.00715

Butterfill, S. A., & Apperly, I. A. (2013). How to Construct a Minimal Theory of Mind. *Mind & Language*, *28*(5), 606–637. https://doi.org/10.1111/mila.12036

Cage, E., Pellicano, E., Shah, P., & Bird, G. (2013). Reputation management: evidence for ability but reduced propensity in autism. *Autism Research: Official Journal of the International Society for Autism Research*, *6*(5), 433–442. https://doi.org/10.1002/aur.1313

Carruthers, P. & Smith, P., K. (1998). *Theories of theories of mind (Part I)*. Cambridge: Cambridge University Press

Colvert, E., Tick, B., McEwen, F., Stewart, C., Curran, S. R., Woodhouse, E., … Bolton, P. (2015). Heritability of Autism Spectrum Disorder in a UK Population-Based Twin Sample. *JAMA Psychiatry*, *72*(5), 1–9. https://doi.org/10.1001/jamapsychiatry.2014.3028

Csibra, G., & Gergely, G. (2006). Social learning and social cognition: The case for pedagogy. In Y. Munakata & M. H. Johnson (Eds.), *Processes of Change in Brain and Cognitive Development. Attention and Performance, XXI* (pp. 249–274). Oxford: Oxford University Press.

Dennett, D. C. (1984). Cognitive Wheels: the Frame Problem of AI. In C. Hookway (Ed.), *Minds, Machines and Evolution*. Cambridge: Cambridge University Press.

Devine, R. T., White, N., Ensor, R., & Hughes, C. (2016). Theory of mind in middle childhood: Longitudinal associations with executive function and social competence. *Developmental Psychology*, *52*(5), 758–771. https://doi.org/10.1037/dev0000105

Dörrenberg, S., Rakoczy, H., & Liszkowski, U. (2018). How (not) to measure infant Theory of Mind: Testing the replicability and validity of four non-verbal measures. *Cognitive Development*, *46*(February 2017), 12–30. https://doi.org/10.1016/j.cogdev.2018.01.001

Dufour, N., Redcay, E., Young, L., Mavros, P. L., Moran, J. M., Triantafyllou, C., … Saxe, R. (2013). Similar Brain Activation during False Belief Tasks in a Large Sample of Adults with and without Autism. *PloS One*, *8*(9), e75468. https://doi.org/10.1371/journal.pone.0075468

Dumontheil, I., Apperly, I. A., & Blakemore, S.-J. (2010). Online usage of theory of mind continues to develop in late adolescence. *Developmental Science*, *13*(2), 331–338. https://doi.org/10.1111/j.1467-7687.2009.00888.x

Frith, C. D., & Frith, U. (2008). Implicit and explicit processes in social cognition. *Neuron*, *60*(3), 503–510. https://doi.org/10.1016/j.neuron.2008.10.032

Godfrey-Smith, P. (2005). Folk psychology as a model. *Philosophers' Imprint*, *5*(6), 1–16. Retrieved from https://www.petergodfreysmith.com/FPasModel-PGS-Imprint.pdf

Gopnik, A., & Wellman, H. M. (1992). Why the Child's Theory of Mind Really is a Theory. *Mind & Language*, *7*(1), 145–171.

Gray, H. M., Gray, K., & Wegner, D. M. (2007). Dimensions of Mind Perception. *Science (New York, N.Y.)*, *315*(5812), 619. https://doi.org/10.1126/science.1134475

Grosse Wiesmann, C., Schreiber, J., Singer, T., Steinbeis, N., & Friederici, A. D. (2017). White matter maturation is associated with the emergence of Theory of Mind in early childhood. *Nature Communications*, *8*(14692). https://doi.org/10.1038/ncomms14692

Happé, F. (1995). The Role of Age and Verbal Ability in the Theory of Mind Task Performance of Subjects with Autism. *Child Development*, *66*(3), 843–855.

Happé, F., Ronald, A., & Plomin, R. (2006). Time to give up on a single explanation for autism. *Nature Neuroscience*, *9*(10), 1218–1220. https://doi.org/10.1038/nn1770

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.

Happé, F. G, Cook, J. L., & Bird, G. (2017). The Structure of Social Cognition : In(ter)dependence of Sociocognitive Processes. *Annu. Rev. Psychol*, *68*(September 2016), 1–25. https://doi.org/10.1146/annurev-psych-010416-044046

Heyes, C. (2014a). Submentalizing: I Am Not Really Reading Your Mind. *Perspectives on Psychological Science*, *9*(2), 131–143. https://doi.org/10.1177/1745691613518076

Heyes, C. (2014b). False belief in infancy: a fresh look. *Developmental Science*, *17*(5), 647–659. https://doi.org/10.1111/desc.12148

Heyes, C. M., & Frith, C. D. (2014). The cultural evolution of mind reading. *Science*, *344*(6190), 1243091. https://doi.org/10.1126/science.1243091

Heyes, C. (2015). Animal mindreading: what's the problem? *Psychonomic Bulletin and Review*, *22*(2), 313–327. https://doi.org/10.3758/s13423-014-0704-4

Hughes, C., & Cutting, A. L. (1999). Nature, Nurture, and Individual Differences in Early Understanding of Mind. *Psychological Science*, *10*(5), 429–432. https://doi.org/10.1111/1467-9280.00181

Hughes, Claire, Jaffee, S. R., Happé, F., Taylor, A., Caspi, A., & Moffitt, T. E. (2005). Origins of individual differences in theory of mind: from nature to nurture? *Child Development*, *76*(2), 356–370. https://doi.org/10.1111/j.1467-8624.2005.00850.x

Krupenye, C., Kano, F., Hirata, S., Call, J., & Tomasello, M. (2016). Great apes anticipate that other individuals will act according to false beliefs. *Science*, *354*(6308), 110–114. https://doi.org/10.1126/science.aaf8110

Kulke, L., Johannsen, J., & Rakoczy, H. (2019). Why can some implicit Theory of Mind tasks be replicated and others cannot? A test of mentalizing versus submentalizing accounts. *PLoS ONE*, *14*(3). https://doi.org/10.1371/journal.pone.0213772

Kulke, L., & Rakoczy, H. (2018). Implicit Theory of Mind – An overview of current replications and non-replications. *Data in Brief*, *16*, 101–104. https://doi.org/10.1016/j.dib.2017.11.016

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

Kulke, L., von Duhn, B., Schneider, D., & Rakoczy, H. (2018). Is Implicit Theory of Mind a Real and Robust Phenomenon? Results From a Systematic Replication Study. *Psychological Science*, *29*(6), 888–900. https://doi.org/10.1177/0956797617747090

Lake, B. M., Ullman, T. D., Tenenbaum, J. B., & Gershman, S. J. (2017). Building Machines That Learn and Think Like People. *Behavioral and Brain Sciences*, *40*(E253), 1–72. https://doi.org/10.1017/S0140525X16001837

Lalonde, C. E., & Chandler, M. J. (2002). Children's understanding of interpretation. *New Ideas in Psychology*, *20*(2–3), 163–198. https://doi.org/10.1016/S0732-118X(02)00007-7

Lecce, S., Bianco, F., Devine, R. T., & Hughes, C. (2017). Relations between theory of mind and executive function in middle childhood: A short-term longitudinal study. *Journal of Experimental Child Psychology*, *163*, 69–86. https://doi.org/10.1016/j.jecp.2017.06.011

Leslie, A. M. (1987). Pretense and Representation: The Origins of "Theory of Mind." *Psychological Review*, *94*(4), 412–426. https://doi.org/10.1037/0033-295X.94.4.412

Leslie, A. M. (1992). Pretense, Autism, and the Theory-of-Mind Module. *Current Directions in Psychological Science*, *1*(1), 18–21. https://doi.org/10.1111/1467-8721.ep10767818

Lillard, A. (1998). Ethnopsychologies: Cultural Variations in Theories of Mind,. *Psychological Bulletin*, *123*(1), 3–32. https://doi.org/10.1037/0033-2909.123.1.3

Livingston, L. A., & Happé, F. (2017). Conceptualising compensation in neurodevelopmental disorders: Reflections from autism spectrum disorder. *Neuroscience and Biobehavioral Reviews*, *80*(June), 729–742. https://doi.org/10.1016/j.neubiorev.2017.06.005

Maibom, H. (2003). The mindreader and the scientist. *Mind and Language*, *18*(3), 296–315. https://doi.org/10.1111/1468-0017.00229

Mayer, A., & Träuble, B. E. (2013). Synchrony in the onset of mental state understanding across cultures? A study among children in Samoa. *International*

Journal of Behavioral Development, *37*(1), 21–28. https://doi.org/10.1177/0165025412454030

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. (2005). The false dichotomy between simulation and theory-theory: the argument's error. *Trends in Cognitve Sciences, 9*(8), 363–364. https://doi.org/10.1016/j.tics.2005.06.003

Mitchell, J. P. (2009). Inferences about mental states. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *364*(1521), 1309–1316. https://doi.org/10.1098/rstb.2008.0318

Moran, J. M. (2013). Lifespan development: The effects of typical aging on theory of mind. *Behavioural Brain Research*, *237*(1), 32–40. https://doi.org/10.1016/j.bbr.2012.09.020

Murray, K., Johnston, K., Cunnane, H., Kerr, C., Spain, D., Gillan, N., … Happé, F. (2017). A new test of advanced theory of mind: The "Strange Stories Film Task" captures social processing differences in adults with autism spectrum disorders. *Autism Research*, *10*(6), 1120–1132. https://doi.org/10.1002/aur.1744

Nelson, K. (2005). Language Pathways into the Community of Minds. In J.W. Astingston & J.A. Barid (Eds.), *Why Language Matter for Theory of Mind* (pp. 26-50). New York: Oxford University Press

Oakley, B. F. M., Brewer, R., Bird, G., & Catmur, C. (2016). 'Theory of Mind' is not

Theory of Emotion: A cautionary note on the Reading the Mind in the Eyes Test. *Journal of Abnormal Psychology*, *125*(6), 818–823. https://doi.org/10.1037/abn0000182

Phillips, J., Ong, D. C., Surtees, A. D. R., Xin, Y., Williams, S., Saxe, R., & Frank, M. C. (2010). A second look at automatic theory of mind: Reconsidering Kovács, Téglás, and Endress (2010). *Psychological Science, 26* (9), 1353-67. https://doi.org/10.1177/0956797614558717

Plomin, R., DeFries, J. C., Knopik, V. S., & Neiderhiser, J. M. (2013). *Behavioral Genetics* (6th Ed.). New York: Worth Publishers.

Powell, L. J., Hobbs, K., Bardis, A., Carey, S., & Saxe, R. (2018). Replications of implicit theory of mind tasks with varying representational demands. *Cognitive Development*, *46*(September 2017), 40–50. https://doi.org/10.1016/j.cogdev.2017.10.004

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

Qureshi, A. W., Apperly, I. A., & Samson, D. (2010). Executive function is necessary for perspective selection, not Level-1 visual perspective calculation: Evidence from a dual-task study of adults. *Cognition*, *117*(2), 230–236. https://doi.org/10.1016/j.cognition.2010.08.003

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

Richardson, H., Lisandrelli, G., Riobueno-Naylor, A., & Saxe, R. (2018). Development

of the social brain from age three to twelve years. *Nature Communications*, *9*(1), 1–12. https://doi.org/10.1038/s41467-018-03399-2

Ronald, A., Happé, F., Hughes, C., & Plomin, R. (2005). Nice and Nasty Theory of Mind in Preschool Children: Nature and Nurture. *Social Development*, *14*(4), 664–684.

Ronald, A., Viding, E., Happé, F., & Plomin, R. (2006). Individual differences in theory of mind ability in middle childhood and links with verbal ability and autistic traits: a twin study. *Social Neuroscience*, *1*(3–4), 412–425. https://doi.org/10.1080/17470910601068088

Rubio-Fernández, P., & Geurts, B. (2013). How to pass the false-belief task before your fourth birthday. *Psychological Science*, *24*(1), 27–33. https://doi.org/10.1177/0956797612447819

Ruffman, T., Garnham, W., Import, A., & Connolly, D. (2001). Does eye gaze indicate implicit knowledge of false belief? Charting transitions in knowledge. *Journal of Experimental Child Psychology*, *80*, 201–224. https://doi.org/10.1006/jecp.2001.2633

Russell, T. A., Schmidt, U., Doherty, L., Young, V., & Tchanturia, K. (2009). Aspects of social cognition in anorexia nervosa: Affective and cognitive theory of mind. *Psychiatry Research*, *168*(3), 181–185. https://doi.org/10.1016/j.psychres.2008.10.028

Samson, D., Apperly, I. A., Braithwaite, J. J., Andrews, B. J., & Bodley Scott, S. E. (2010). Seeing it their way: evidence for rapid and involuntary computation of what other people see. *Journal of Experimental Psychology. Human Perception and Performance*, *36*(5), 1255–1266. https://doi.org/10.1037/a0018729

Saxe, R. (2005). Against simulation: the argument from error. *Trends in Cognitive Sciences*, *9*(4), 174–179. https://doi.org/10.1016/j.tics.2005.01.012

Schaafsma, S. M., Pfaff, D. W., Spunt, R. P., & Adolphs, R. (2015). Deconstructing and reconstructing theory of mind. *Trends in Cognitive Sciences*, *19*(2), 65–72. https://doi.org/10.1016/j.tics.2014.11.007

Schurz, M., Radua, J., Aichhorn, M., Richlan, F., & Perner, J. (2014). Fractionating theory of mind: A meta-analysis of functional brain imaging studies. *Neuroscience and Biobehavioral Reviews*, *42*, 9–34. https://doi.org/10.1016/j.neubiorev.2014.01.009

Senju, A., Southgate, V., Snape, C., Leonard, M., & Csibra, G. (2011). Do 18-month-olds really attribute mental states to others? A critical test. *Psychological Science*, *22*(7), 878–880. https://doi.org/10.1177/0956797611411584

Shahaeian, A., Peterson, C. C., Slaughter, V., & Wellman, H. M. (2011). Culture and the Sequence of Steps in Theory of Mind Development. *Developmental Psychology*, *47*(5), 1239–1247. https://doi.org/10.1037/a0023899

Slaughter, V., & Perez-Zapata, D. (2014). Cultural Variations in the Development of Mind Reading. *Child Development Perspectives*, *8*(4), 237–241. https://doi.org/10.1111/cdep.12091

Spunt, R. P., & Adolphs, R. (2015). Folk Explanations of Behavior: A Specialized Use of a Domain-General Mechanism. *Psychological Science*, *26*(6), 724–736. https://doi.org/10.1177/0956797615569002

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. https://doi.org/10.1073/pnas.1511905112

Wang, L., & Leslie, A. M. (2016). Is Implicit Theory of Mind the "Real Deal"? The Own-Belief/True-Belief Default in Adults and Young Preschoolers. *Mind and Language*, *31*(2), 147–176. https://doi.org/10.1111/mila.12099

Warnell, K. R., & Redcay, E. (2019). Minimal coherence among varied theory of mind measures in childhood and adulthood. *Cognition*, *191*(June), 103997. https://doi.org/10.1016/j.cognition.2019.06.009

Wellman, H. M. (2014). *Making Minds: How Theory of Mind Develops*. (P. Bloom & S. A. Gelman, Eds.). New York, NY: Oxford University Press.

Wellman, H. M., Cross, D., & Watson, J. (2001). Meta-Analysis of Theory-of-Mind Development: The Truth about False Belief. *Child Development*, *72*(3), 655–684. https://doi.org/10.1111/1467-8624.00304

Wellman, H. M., & Liu, D. (2004). Scaling of Theory-of-Mind Tasks. *Child Development*, *75*(2), 523–541. https://doi.org/10.1111/j.1467-8624.2004.00691.x

White, S., Hill, E., Happé, F., & Frith, U. (2009). Revisiting the strange stories: Revealing mentalizing impairments in autism. *Child Development*, *80*(4), 1097–1117. https://doi.org/10.1111/j.1467-8624.2009.01319.x

Willard, A. K., & McNamara, R. A. (2019). The Minds of God(s) and Humans: Differences in Mind Perception in Fiji and North America. *Cognitive Science*, *43*(1), 1–30. https://doi.org/10.1111/cogs.12703

2. Understanding Individual Differences in Theory of Mind via Representation of Minds, Not Mental States

This chapter is presented as a published article and is an exact copy of the following journal publication:

Conway, J.R., Catmur, C. & Bird, G. (2019). Understanding Individual Differences in Theory of Mind via Representation of Minds, Not Mental States. *Psychonomic Bulletin & Review*, *26*(3), 798-812. https://doi.org/10.3758/s13423-018-1559-x

Corresponding author:

J.R. Conway, Social, Genetic and Developmental Psychiatry Centre, Institute of Psychiatry, Psychology and Neuroscience, King's College London, London SE5 8AF, United Kingdom.

jane_rebecca.conway@kcl.ac.uk

Psychonomic Bulletin & Review
https://doi.org/10.3758/s13423-018-1559-x

THEORETICAL REVIEW

Understanding individual differences in theory of mind via representation of minds, not mental states

Jane R. Conway[1,2] · Caroline Catmur[3] · Geoffrey Bird[1,2]

Abstract
The human ability to make inferences about the minds of conspecifics is remarkable. The majority of work in this area focuses on mental state representation ('theory of mind'), but has had limited success in explaining individual differences in this ability, and is characterized by the lack of a theoretical framework that can account for the effect of variability in the population of minds to which individuals are exposed. We draw analogies between faces and minds as complex social stimuli, and suggest that theoretical and empirical progress on understanding the mechanisms underlying mind representation can be achieved by adopting a 'Mind-space' framework; that minds, like faces, are represented within a multidimensional psychological space. This Mind-space framework can accommodate the representation of whole cognitive systems, and may help to explain individual differences in the consistency and accuracy with which the mental states of others are inferred. Mind-space may also have relevance for understanding human development, intergroup relations, and the atypical social cognition seen in several clinical conditions.

Keywords Theory of mind · Face-space · Individual differences · Social cognition · Mind-space

Introduction

Minds, like faces, are a special set of stimuli in the social environment. They are a dynamic source of information about the behavior of conspecifics, with relevance for many aspects of everyday life, from the enjoyment of friendships to how a jury assesses the accused. Understanding how we represent the minds of other humans is therefore a particularly important aim. For the past 27 years, the idea that faces are represented within a multidimensional psychological space has provided a unifying theoretical framework that explains multiple experimental effects and informs new predictions (Valentine, 1991; Valentine, Lewis, & Hills, 2016). The concept of 'Face-space' has brought coherence to a large literature, and offers a psychological model of how these multifarious stimuli are processed. In contrast to the literature on face processing, the study of how minds are represented lacks a coherent organizational framework (Happé, Cook, & Bird, 2017).

We suggest that the study of mind representation would benefit from the adoption of a 'Mind-space' framework – where minds are represented within a multidimensional space – in the same way as the face processing literature has from the introduction of Face-space (Oosterhof & Todorov, 2008; Todorov, Olivola, Dotsch, & Mende-Siedlecki, 2015; Todorov, Said, Engell, & Oosterhof, 2008; Valentine et al., 2016). We argue that adopting the Mind-space framework would allow explanation of individual differences in the ability to represent minds, and also in the ability to infer mental states. Here, we use the term 'mind' to refer to an individual's complete set of cognitive systems, and the term 'mental state' to refer to the representational content generated by that set of systems. The probability of specific mental states is dependent on the properties of the mind to which they are ascribed. Therefore, understanding individual differences in the representation of minds allows individual differences in the accuracy of mental state inference to be explained. For example, the mental state 'Everyone in the world loves me' would be more likely to be generated by a mind that has the property of

✉ Jane R. Conway
jane_rebecca.conway@kcl.ac.uk

[1] MRC Social, Genetic and Developmental Psychiatry Centre, Institute of Psychiatry, Psychology and Neuroscience, King's College London, London SE5 8AF, UK

[2] Department of Experimental Psychology, University of Oxford, Oxford OX1 4AL, UK

[3] Department of Psychology, Institute of Psychiatry, Psychology & Neuroscience, King's College London, London SE1 1UL, UK

Published online: 16 January 2019

a high degree of narcissism, than one without such a property. Therefore, people who are better able to characterize the specific mind generating a mental state are likely to be more accurate at inferring that mental state. Accordingly, this paper proposes a mechanism by which the ability to represent minds in Mind-space explains skill in accurately inferring mental states.

We outline how the Mind-space framework can enable the following necessary advances: Describe how people represent all properties of minds; explain variance in the quality and structure of such representations; elucidate the processes by which another's mental states are inferred; and explain individual differences in the accuracy of mental state inference. In order to do so we will make three independent, but related, arguments, namely:

1. Understanding individual differences in representation of mental states is difficult within current frameworks.
2. Although mental states are a product of the individual mind that gave rise to them, representation of minds is largely absent from empirical and theoretical work on mental state inference.
3. Adoption of a Mind-space framework is one way in which representation of minds can be incorporated into the process of mental state inference, and in doing so one can better understand individual differences in mental state inference.

Understanding individual differences in theory of mind

To date, the study of understanding other minds has focused on how people represent others' mental states, such as thoughts and beliefs; this ability is most often termed 'theory of mind' (Baron-Cohen, Leslie, & Frith, 1985; Premack & Woodruff, 1978). Despite the thousands of studies referencing theory of mind, it is still unclear what individual differences in the ability represent (Bird, 2017; Bartsch & Estes, 1996; Conway & Bird, 2018). This may be due to the lack of theories addressing the underlying psychological processes involved in the representation of mental states (Baker, Jara-Ettinger, Saxe, & Tenenbaum, 2017; Schaafsma, Pfaff, Spunt, & Adolphs, 2015; Spunt & Adolphs, 2015), and how the contents of such representations are derived. Therefore, explanations for individual differences in theory of mind have been limited to invoking domain-general inferential processes such as language (Milligan, Astington, & Dack, 2014) or executive function (Carlson & Moses, 2001; Devine & Hughes, 2014; Hughes, 1998), rather than domain-specific representational structures. Although it is clear that variance in domain-general processes may influence performance on theory of mind tests, variance within these domains would influence performance on most tasks, and variance in such domain-general processes does not inform what it is to be better or worse specifically at representing mental states, and why (Conway & Bird, 2018; Bird, 2017).

Understanding individual differences in theory of mind would be aided by a model of what determines the difficulty of representing different types of mental states *within* an individual. Surprisingly, although there is considerable debate in the literature as to what qualifies as a mental state – for example whether someone's visual perspective (Samson, Apperly, Braithwaite, Andrews, & Bodley Scott, 2010) or emotional state (Oakley, Brewer, Bird, & Catmur, 2016) qualifies as a mental state, or whether the term should be reserved for representation of propositional attitudes (Butterfill & Apperly, 2013; Leslie, 1987) – there is considerable agreement that certain types of mental state are harder to represent than others. For example, few experts would disagree that it is harder to represent false beliefs (beliefs held by an individual that you know to conflict with reality) than true beliefs (Leslie, 1987; Wimmer & Perner, 1983). Despite this agreement, however, as far as we are aware there is little understanding of what makes some mental states harder to represent than others, beyond the fact that representation of some types of mental state makes greater demands on domain-general processes such as working memory, language, or executive function, than representation of other types of mental state.

In the absence of such understanding, it is important to understand the basis for the consensus of opinion as to the relative difficulty of representing different types of mental state. One important influence is the work of Wellman and colleagues (Shahaeian, Peterson, Slaughter, & Wellman, 2011; Wellman, Fang, Liu, Zhu, & Liu, 2006; Wellman, Fang, & Peterson, 2011; Wellman & Liu, 2004) within the field of developmental psychology. This work has described the developmental trajectory of mental state understanding and noted that understanding of certain types of mental state tends to occur earlier in development than understanding of other types of mental state (e.g., understanding of desires occurs before understanding of beliefs). Such evidence has been used to support the idea that certain types of mental state are more difficult to represent than others. However, the order in which different types of mental state are understood varies across cultures, for instance children in Iran and China tend to understand the relationship between seeing and knowing before appreciating that people can have diverse beliefs, whereas the reverse order is observed in children from Australia and the USA (Shahaeian, Nielsen, Peterson, & Slaughter, 2014a; Shahaeian, Nielsen, Peterson, Aboutalebi, & Slaughter, 2014b; Shahaeian et al., 2011; Slaughter & Perez-Zapata, 2014; Wellman et al., 2006, 2011). This makes it likely that the order in which children understand different types of mental state may instead depend on environmental

factors such as when they are taught about each type of mental state (Heyes & Frith, 2014), rather than providing any explanation of, or justification for, differential difficulty of mental state representation (Conway & Bird, 2018; Bird, 2017). Moreover, it is also possible that the age at which children can represent different types of mental state is governed by the degree to which they recruit domain-general processes of executive function or language, and the developmental timetable of these processes (Devine & Hughes, 2014; Milligan et al., 2014; Sabbagh, Xu, Carlson, Moses, & Lee, 2006).

An absence of minds in tests of theory of mind

Theory of mind is typically defined as the ability to *represent* mental states. In contrast, theory of mind measures tend to test the ability to make accurate mental state *inferences*. This distinction is important; on any particular test one could make an inaccurate mental state inference yet still represent a mental state. In such a situation there is no deficit in the representation of mental states, but rather a deficit in accurately inferring the content of a particular mental state.

Theory of mind tests tend to require the participant to infer the mental state of a protagonist in a certain situation (Baron-Cohen et al., 1985; Dziobek et al., 2006). The 'correct' mental state inference is typically determined by the test authors based on rational consensus. Such an operationalization results in a binary response measure: one either can, or cannot, make the correct mental state inference. As a consequence, these measures are not sensitive to subtle variance in the quality of mental state inference processes, and ignore perhaps the most important source of inferential error: representation of the mind giving rise to the mental state.

Specifically, existing tests of mental state inference largely fail to take account of the variability in the populations of minds available for representation, and the degree to which this variability is incorporated into mental state inference. An individual is exposed to many different minds, and 'mind type' – the collection of long- and short-term attributes characterizing a particular mind – is likely to influence the kind of mental states a particular mind produces. One can easily imagine that, even in the same objective situation, an optimistic mind may produce very different mental states from a pessimistic mind; an autistic mind different mental states from a neurotypical mind; and an adult mind different mental states from a child's mind. This variance in mental states as a function of mind type – a crucial component of the accuracy of naturalistic mental state inference – is absent from tests of theory of mind that make use of an anonymous protagonist about whom nothing is known. Even those tests that introduce well-formed characters with distinct personalities, tests that have the potential to examine the degree to which mental state inference varies as a function of the protagonist's mind type, do not explicitly score this aspect of mental state inference (Dziobek et al., 2006).

Furthermore, although the majority of tests of theory of mind have examined the representation and inference of mental states – the *content* of someone's mind – there are also multiple *processes* of mind available for representation. The degree to which these are represented, and the accuracy of their representation, is likely to contribute to variance in the accuracy of mental state inference. Several of these mental processes have been addressed by cognitive science, such as memory, attention, and spatial reasoning; but the degree to which they are represented as properties of others' minds has been less well studied (Camerer, Ho, & Chong, 2004; Coricelli & Nagel, 2009). Moreover, such work has rarely been linked to the representation of other aspects of mind. It is strange that, for example, the evaluation of others' working memory or metacognitive ability is not linked theoretically to representing their mental states (e.g., thoughts and beliefs), when both constitute properties of another's mind that are available for representation and which may help predict their subsequent behavior. These processes can be described as features of minds in the same way as personality traits such as optimism or aggressiveness, and may also produce variance in mental states despite an identical situation. A forgetful mind may give rise to different mental states than a mind with good memory; a more intelligent mind may give rise to different mental states than a less intelligent one; and so on. The degree to which individuals incorporate such information in their inference of mental states is also largely untested in current tests of mental state inference.

Without a theoretical framework that addresses variance in other minds and their representation, explanations of individual differences in theory of mind will remain limited to domain-general abilities, rather than the quality of domain-specific representational content and the inferential processes specific to accurate mental state representation. We argue that the development of a theoretical framework that describes representation of whole cognitive systems, i.e. of minds in their entirety, would contribute to the understanding of those psychological processes giving rise to more or less accurate inference of another's mental states.

Mind-space: A new framework for understanding the representation of minds

We suggest that theoretical and empirical progress on understanding mind representation, and separately the inference of mental states, can be achieved by adopting a Mind-space framework; that minds, like faces, are represented within a multidimensional psychological space (Fig. 1). The Face-space framework was motivated by the lack of a theory that

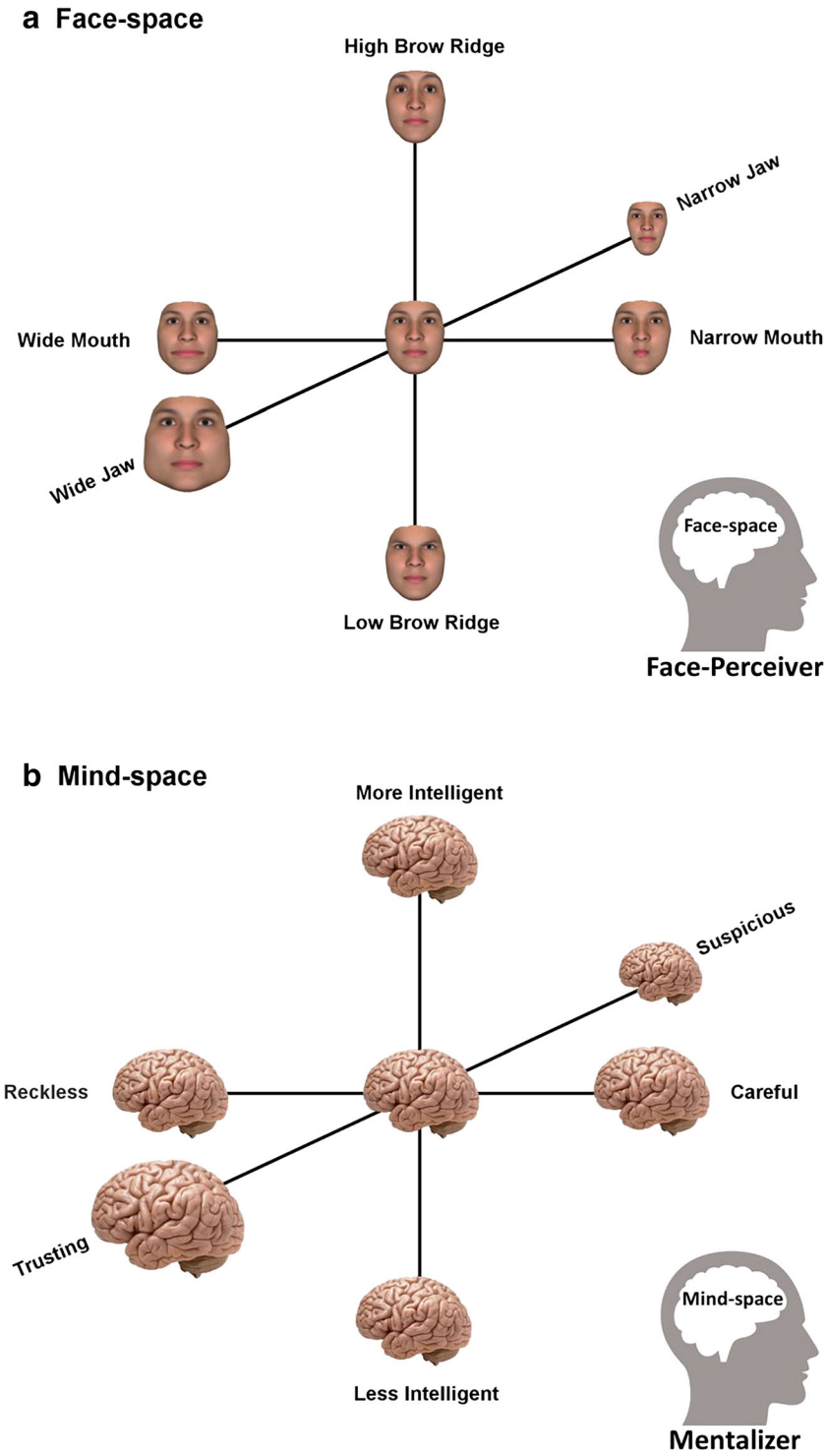

Fig. 1 Multidimensional representational spaces: Face-space and Mind-space. In this example of Face-space (**A**), faces are represented on three orthogonal dimensions of brow ridge height, jaw width, and mouth width. In this Mind-space example (**B**), minds are represented on orthogonal dimensions which allow them to be individuated from one another. Dimensions may reflect cognitive abilities (e.g., intelligence), behavioral tendencies (e.g., recklessness), or personality traits (e.g., suspiciousness). (The human brain image is reproduced with permission from Dan Heighton).

could account for seemingly disparate findings in the face-processing literature, and by the need for a model that would reflect the effect of variance in faces experienced by the individual (Valentine, 1991; Valentine et al., 2016). Face-space is a multidimensional space, the dimensions of which are unspecified but can represent any discriminable aspect of

faces, from structural aspects such as nose length to more abstract traits, like attractiveness or trustworthiness (Fig. 1a). In someone's Face-space, every individual face is represented as a vector along multiple dimensions; the population of experienced faces is normally distributed and the intercept of the axes reflects the dimensional means (Valentine et al., 2016). Although the idea that representations of stimuli are structured along dimensions extends to most percepts, including features of non-social objects such as color, size, or tilt (Thompson & Burr, 2009), Face-space has provided a psychological model to explain a range of empirical findings and acts as a unifying theory of how representations of such complex social stimuli may be structured. Effects explained by the Face-space framework include: why distinctive faces are better recognized than typical faces, even when inverted (Valentine, 1991); why there is an own-ethnicity face recognition bias (Chiroro & Valentine, 1995); perceptual adaptation effects (Jeffery & Rhodes, 2011; Jiang, Blanz, & O'Toole, 2006, 2009; Leopold, O'Toole, Vetter, & Blanz, 2001; Rhodes, Jeffery, Watson, Clifford, & Nakayama, 2003; Webster, Kaping, Mizokami, & Duhamel, 2004); and why children's face-processing abilities may differ to adults' (de Heering, Rossion, & Maurer, 2012; Hills, Holland, & Lewis, 2010).

We suggest that a Mind-space framework can overcome current theoretical limitations on mind representation. In common with faces, minds present many dimensions on which they may be similar to, or discriminated from, one another. It is therefore possible to represent individual minds within a multidimensional space, analogous to how faces are represented within Face-space (Fig. 1b). There is no requirement for the axes that represent the space to be orthogonal, meaning that the space can be constructed such that the relationship between axes represents the co-variance between properties of minds encountered in the real world. For example, if a bivariate correlation exists such that one property of minds, suspiciousness, predicts another property, such as aggressiveness, then axes can be constructed such that movement along the suspiciousness dimension causes movement along the aggressiveness dimension. Within such a Mind-space framework, an individual's representation of another mind can be described as a single vector, or location, in a space determined by multiple axes.

Representation of the whole cognitive system and variability in mind type

The Mind-space framework allows multiple aspects of mind to be represented within one model; one dimension may represent suspiciousness, another working-memory ability, and another political persuasion. However, this is only necessary if people actually represent those properties of minds that allow them to be differentiated, in addition to the contents of their mental states. Evidence for such representation is provided by examples of 'recipient design' – the adaptation of one's communications to better suit a specific addressee (Blokpoel et al., 2012). For example, several studies using the Tacit Communication Game (Stolk, Noordzij, Verhagen, et al., 2014a; Stolk, Noordzij, Volman, et al., 2014b) demonstrated that communicators modulate their communicative behavior as a function of whether they think they are communicating with someone younger than them (Newman-Norlund et al., 2009; Stolk, Hunnius, Bekkering, & Toni, 2013). The adaptations made by communicators are frequently attributed to the representation of the addressee's mental states, e.g., beliefs or knowledge (Blokpoel et al., 2012; Newman-Norlund et al., 2009; Stolk, Noordzij, Verhagen, et al., 2014a; Stolk, Noordzij, Volman, et al., 2014b; Stolk et al., 2013). However, modulation of communicative behavior as a function of addressee age suggests that communicators are representing the cognitive processes of the addressee (such as their working memory capacity or inspection time) in addition to their mental states. Similarly, the adoption of 'elderspeak' when communicating with older adults, by using slower, shorter sentences (Kemper & Harden, 1999; Williams, Kemper, & Hummert, 1995), likely reflects representations of the memory and processing speed of older adults. Indeed, accurate comprehension of others' communications can be affected by representations of their linguistic background. The 'speaker-model' account of word recognition suggests that listeners disambiguate words with different dominant meanings in British compared to American English by first identifying the speaker's dialect and then adopting that model for subsequent interpretations (Cai et al., 2017).

Neuroimaging studies have suggested that the medial prefrontal cortex (mPFC), a brain region in the 'theory of mind network', may encode information about other people and their personality traits (Hassabis et al., 2014; Heleven & Van Overwalle, 2015, 2018). Suppression effects in the ventral mPFC have been observed with repetition of the same trait (Ma et al., 2014) or person (Heleven & Van Overwalle, 2015). Ma et al. (2014) found suppression effects both for pairs of stimuli that signified the same trait (e.g., honesty + honesty) and for pairs that signified the opposite trait (e.g., dishonesty + honesty). This latter finding holds particular significance for the Mind-space theory, as it implies that traits of others' minds are represented along dimensions and not categorically (Heleven & Van Overwalle, 2018).

The relevance of Mind-space to theory of mind

Mental states are a product of the minds that give rise to them. Accurate and specific inference of the contents of another's mental states is therefore likely aided by representing multiple features of minds and variability in mind type. For example, theory of mind is commonly tested using a false-belief task such as the Sally-Anne task (Fig. 2, Panel I) (Baron-Cohen

et al., 1985). In this task participants are introduced to two characters, Sally and Anne, and are informed that Sally has a ball that she places into her basket before leaving the room. While Sally is away, Anne takes Sally's ball and places it in her own box. Participants are asked where Sally will look for her ball on her return. This type of paradigm is frequently described as providing the strongest evidence of mental state representation (Baron-Cohen et al., 1985; Dennett, 1978) because successful performance requires the ascription of a false belief: that Sally will act based on a false belief that is inconsistent with where the object actually is and where the participant knows it to be located. Participants are therefore determined to have given a correct answer if they respond that Sally will look in her basket, and an incorrect answer if they respond that Sally will look in Anne's box. While this task is relatively straightforward, one can imagine that what is deemed a correct answer is likely to change if we know that Sally has high levels of suspiciousness and is likely to suspect Anne has stolen her ball. In this case we may imagine that Sally will first look in Anne's box to check her assumption that Anne has stolen her ball. In this scenario, a participant who has a dimension of suspiciousness in their Mind-space and who recognizes that Sally is at the extreme end of this dimension is likely to be more accurate when inferring the content of Sally's mental states than another individual who either does not represent suspiciousness as a property of minds, or who cannot locate Sally accurately along the suspiciousness dimension (Fig. 2).

It can therefore be seen that adopting a multidimensional representational space offers a framework for investigating individual differences in the ability and propensity to represent the properties of other minds, and an explanation of differences in the accuracy and specificity with which the contents of mental states can be inferred. Within the Mind-space framework, the model of a specific other's mind would serve as a function that takes as its input the context the other is in, and outputs the likelihood of particular mental states. In statistical terms, one can represent this as the probability of a particular mental state given a particular context and the position of the target mind within an individual's Mind-space. Individual differences in the representation of other minds, and in the

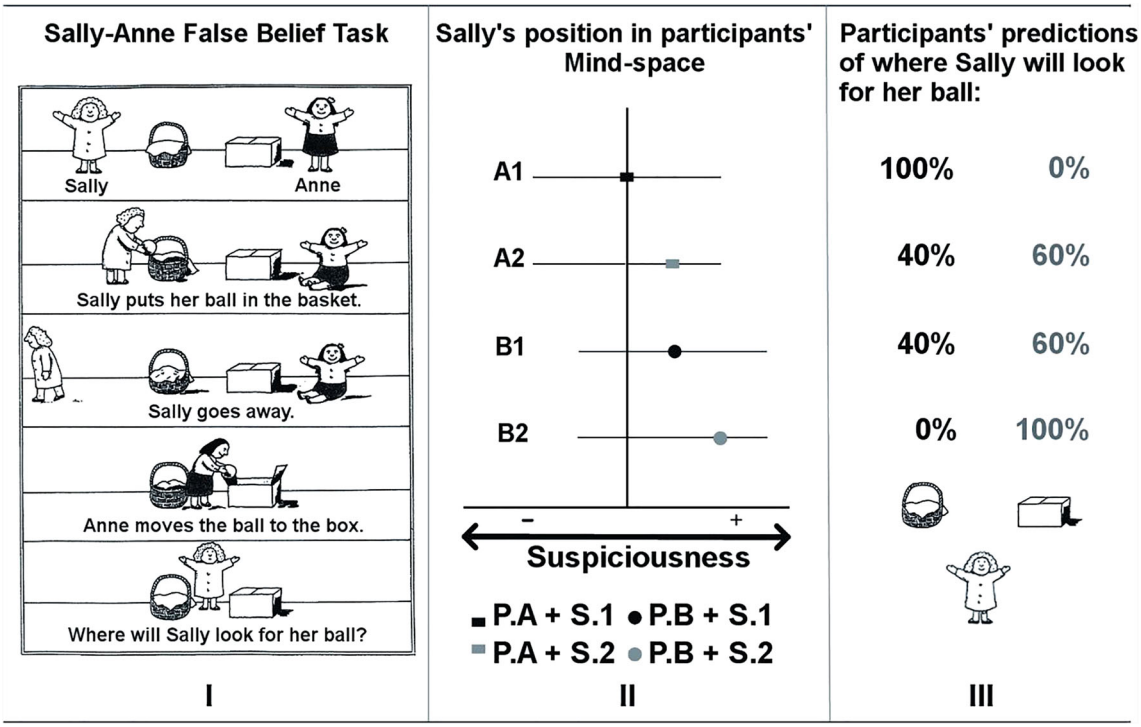

Fig. 2 Suspicious minds: How Mind-space explains performance on the Sally-Anne false belief task. In this test of theory of mind (*Panel I*), to respond correctly participants (**P**) must represent Sally's mental state in the absence of any additional information about her, Anne, or the situation (**S**). In this scenario (Situation 1), an average participant (**P.A; Panel II**) would likely represent Sally at the population mean of suspiciousness in his/her Mind-space, and expect Sally to think that her ball was in the basket where she left it (**Panel III**). The same average participant (**P.A**) in a different situation (**S.2**), having prior knowledge that Sally has high levels of suspiciousness, would represent Sally at a position of high suspiciousness further from the mean. Participant A in Situation 2 might therefore represent Sally as believing that Anne may have moved her ball to the box. Another participant (**P.B**) who has been exposed to an untrustworthy population may, in the absence of any information (**S.1**), have a mean suspiciousness higher than the population average, and, positioning Sally at the mean in his/her Mind-space, similarly represent Sally as believing that Anne may have moved her ball to the box. In Situation 2, having prior knowledge that Sally has high levels of suspiciousness, Participant B would represent Sally further from his/her mean and attribute to Sally the belief that Anne has certainly moved her ball to the box. This example demonstrates how an individual's representation of Mind-space combines with situational information to influence the inference of another person's mental state. (Panel 1 reproduced with permission from Frith, 2003).

accuracy of mental state inference, would therefore be due to one or more of the following factors:

1. Fundamental features of the architecture of an individual's representation of Mind-space such as the complexity of the representational space in terms of the number of dimensions and representation of the co-variance between dimensions, or the 'granularity' or level of detail represented in each dimension.
2. The accuracy with which one can locate a target mind within one's Mind-space on the basis of a sample of behavior.
3. The propensity of an individual to represent minds within Mind-space, and the degree of effort expended in locating a target mind within Mind-space with a high degree of precision.
4. The accuracy of the mapping between position in Mind-space and specific mental states (e.g., the mapping from Panel 2 to Panel 3 in Fig. 2), and the propensity to use position in Mind-space when making a mental state inference.

The self, metacognition, and Mind-space

The question of whose mind is modelled as the default – i.e., the mind that is used to ascertain the probability of particular mental states given situational information only – has long been a topic of debate within the theory of mind literature. One prominent account, the Simulation Theory, posits that one uses one's own mind as this default, to run a simulation that outputs the probability of specific mental states, the most likely of which is then ascribed to the target (Carruthers & Smith, 1996). In this account, egocentric effects are likely to be observed; one attributes the mental state one's own mind would generate if in the same situation as the target. Under the Mind-space framework, however, if one has the propensity to use position in Mind-space when inferring mental states, one does not use one's own mind as a model of others. Rather, one represents a target mind's position in Mind-space, or in the absence of any individuating information (i.e., for an anonymous protagonist), likely assumes the mind to be in the center of Mind-space (representing the population average on each dimension of Mind-space).

The distance between the center of an individual's Mind-space and where they believe their own mind to be located within Mind-space is likely to vary across individuals. Some individuals would judge themselves to be average on some or all dimensions, while others would judge themselves to be more extreme. We use the term 'metacognitive accuracy' to refer to the degree to which an individual can accurately locate themselves in Mind-space; those with high metacognitive accuracy would, for example, be able to judge their IQ relative to the rest of the population, whereas someone with low metacognitive accuracy would either over- or under-estimate their IQ relative to the rest of the population.

The distance between the center of an individual's Mind-space and where they judge their own mind to be in Mind-space is likely to have important implications for how accurately they can infer the mental states of an anonymous target; furthermore, the effect of this distance on the accuracy of mental state inferences will be moderated by the individual's metacognitive accuracy. The privileged access to one's own mental states is likely to result in extensive and enduring mappings between the location one believes oneself to occupy in Mind-space and the mental states experienced in particular situations, due to the fact that one receives more data about one's own mental states than others' mental states, and mappings are likely to be less variable than those provided by experience of a variety of other individuals. Thus, an individual who locates their own mind in the center of their Mind-space can use their own mind as a model for an anonymous target mind (which is most likely to be also in the center of their Mind-space), or for minds they judge to be similar to their own (i.e., estimated also to be in the center of their Mind-space). Accuracy when inferring the mental states of such target minds will therefore depend on two factors: (1) The individual having good metacognitive accuracy and therefore truly being in the center of their Mind-space; and (2) the individual accurately locating targets within Mind-space (and therefore the targets are truly in the center of their Mind-space). Providing these two conditions are satisfied, good accuracy is afforded by the increased accuracy of the mappings between location in Mind-space and the probability of particular mental states resulting from the privileged access the individual has to their own mental states. If an individual has good metacognition but does not locate their own mind at the center of their Mind-space, then their own mind is not a good model for an anonymous target mind (who would be located at the center); however, if they can accurately locate targets within their Mind-space then their own mind will act as a good model for targets similar to the self.

In contrast, if the individual has poor metacognitive accuracy but can accurately locate others in Mind-space, then they are likely to make inaccurate inferences concerning the mental states of targets whom they either believe to have a mind like their own, or targets who actually do have a mind similar to their own. Furthermore, when poor metacognitive accuracy but an intact ability to locate others within Mind-space is combined with accurate mappings between locations in Mind-space and mental states, then the individual would exhibit a decreased ability to predict the likelihood of their own mental states – a situation likely to result in disorders characterised by an atypical sense of self, self-delusions, or a reduced sense of agency.

Non-metacognitive aspects of the self may also impact upon one's Mind-space. For example, an individual very high on trait agreeableness may be less likely to attribute negative attributes to others, or attribute less extreme negative attributes. This would result in a Mind-space where negative attributes are skewed towards low scores, have low mean values or granularity, or co-variances are inaccurately represented. Similarly, individuals who tend to attribute behavior to aspects of the situation rather than the characteristics of the target's mind may be slower to: (1) construct a Mind-space; (2) learn to locate targets within Mind-space in general; or (3) learn to locate a specific target within their Mind-space.

Relationship to existing theories

When considering the relationship between the current proposal and existing theories it is first worth acknowledging what is not novel about the proposal. Most obviously, it is clear that trait models have previously been used in psychology, notably within the field of personality where dominant models suggest that variance in personality can be explained using a model with five or six trait dimensions (Ashton & Lee, 2007; Goldberg, 1990; McCrae, 1989). Of more relevance to Mind-space are existing dimensional models of how we represent individuals, groups, or other agents. For example, Gray, Gray, and Wegner (2007) suggested that judgments regarding other agents' (e.g., children, robots, supernatural beings) ability to feel pain, emotions, have personalities, etc. can be accounted for by a two-dimensional model of whether they are capable of having experiences, and whether they have agency. Perhaps closer to the concept of Mind-space is the work of Fiske and colleagues (Fiske, Cuddy, & Glick, 2007), who have convincingly demonstrated that the dimensions of warmth and competence explain a large degree of the variance in how individuals and groups are perceived. It is therefore clear that the idea that humans can represent other humans (and non-human agents) on trait dimensions which can be described by a reduced set of dimensions or factors is not novel.

The novel feature of the Mind-space proposal is that it explains how variance in *representing minds*; specifically, variance in the structural properties of the multidimensional space within which minds are represented, can explain individual differences in *the ability to make mental state inferences*. In this context, it is important to consider how it relates to the work of Tamir and Thornton (Tamir & Thornton, 2018; Tamir, Thornton, Contreras, & Mitchell, 2016), who have developed an independent proposal relating trait representation to mental states and actions.

Tamir and Thornton's primary aim is not to explain individual differences in the ability to make mental state inferences, but rather to identify the information used to make social predictions and how it is represented. Accordingly, they posit the existence of a multilayered dimensional framework where the layers correspond to others' actions, mental states, and traits, and each of these layers can be characterized on the basis of three dimensions. They put forward an interesting account of how transitions between these layers may allow the prediction of social behavior, an account that is compatible with several existing dimensional theories of person and agent perception (e.g., Fiske et al., 2007; Gray et al., 2007).

As mentioned above, this account does not address individual differences (in the dimensional structure of the multilayered framework, the ability to locate a target mind accurately within it, or the propensity to do so). Furthermore, the nature of the mental state representations is very different in the Tamir and Thornton and Mind-space frameworks. To illustrate, the dimensions used to represent mental states in the Tamir and Thornton framework are rationality, social impact, and valence; and these can be used to encode concepts such as emotions (disgust) and states of mind (intoxicated, weary, fatigued), or to distinguish between mental state types (opinion, belief, thought). Under the Mind-space framework, however, it is minds, not mental states, that are represented dimensionally. Mental *states* are *not* represented dimensionally because the Mind-space framework attempts to explain variance in the ability to infer the content of specific mental states, and in many cases this content is unlikely to be represented in a dimensional structure. For instance, in the case of the Sally-Anne example (Fig. 2), propositional attitudes such as 'John believes that Sally will look for her ball in her basket' and 'John believes that Sally will look for her ball in Anne's box' are very different, yet presumably would be located in exactly the same location in the Tamir and Thornton framework, as that framework distinguishes between mental state types (e.g., 'belief' vs. 'desire'), but does not encode mental state content.

Correct inference of specific mental state representations relies on consideration of situational factors, which are currently outside the Tamir and Thornton framework. However, recognition of the importance of situational factors prompts consideration of how the hypothesized role for Mind-space in the inference of mental states can be reconciled with recently developed computational models of mental state inference that describe how mental states might be predicted on the basis of the situation. We suggest that the addition of Mind-space terms to these computational models of mental state inference may significantly improve their predictive validity, and allow them to be tailored to specific individuals or groups.

An example of such a model is the Bayesian Theory of Mind (BToM) model of Baker et al. (2017), which models the computational basis of 'core mentalizing': meta-representation of the percepts, desires, and beliefs of a rational agent inferred from their actions in a given physical spatial environment. In the BToM framework, it is assumed that the agent updates its beliefs based on percepts and prior

knowledge, and acts rationally to achieve its desires with maximum efficiency and minimum cost. Inference of the agent's beliefs and desires is achieved through inversion of a generative model which describes how mental states cause actions. The generative model is conditioned on observed actions, and representation of unobserved mental states (percepts, beliefs, desires) is thought to be a result of Bayesian inference. The BToM model has been shown to be a successful model of human mental state inference (at least in constrained environments with a limited set of possible desires and beliefs, Baker et al., 2017). However, although BToM is a successful model of how such inference may work in general, by incorporating the position of a specific agent in a particular individual's Mind-space one can further constrain the set of inferences likely to be made about the agent's mental states by that individual (Jacob & Jeannerod, 2005). Furthermore, one can explain why that individual's inference differs from that of another individual, and therefore why one individual is more or less accurate than another. Inclusion of the Mind-space framework within Bayesian generative models of mental state inference may therefore increase their specificity with respect to particular individuals. In addition to increased specificity, modelling of an agent's position within an individual's Mind-space, particularly on dimensions such as intelligence, attention to detail, and perseveration, is likely to explain the degree to which the individual expects the agent to update the content of its mental states as a function of experience within a dynamic system.

For example, the probability of a particular mentalizer inferring that an individual target mind holds a certain mental state is a function of the prior probability of:

- that mental state in general;
- the probability of the mental state conditional upon the situation the target is in;
- and the position of the target in the mentalizer's Mind-space.

The relative influence of situational factors and the target's position in the mentalizer's Mind-space on the posterior estimate of the probability of the target's mental state will be determined by the precision of the prediction each affords. For example, if the target is being chased by a bear then one may make a very precise prediction as to their mental state on the basis of the situation they are in, whereas the prediction based on their position in the mentalizer's Mind-space is likely to be less precise. In this situation, the posterior prediction of the target's mental state will be governed more by the context than by their position in the mentalizer's Mind-space. There may be other contexts where the situation allows a less precise prediction of the target's mental state, and position in Mind-space a more precise prediction. In this case, the mentalizer's posterior prediction would be based more on the target's position in the mentalizer's Mind-space than the situation the target is in. Note however, that even if it is the case that position in Mind-space affords a precise prediction of the relevant mental state in principle, it may still be the case that the mentalizer has an imprecise representation of the location of the target in their own Mind-space. As a consequence, the prediction of the probability of a certain mental state given a target's position in Mind-space will also be imprecise (see Fig. 3).

Predictions and implications of the Mind-space framework

The development of Face-space is thought to be experience-dependent. The space is optimized for the population of faces to which one has been exposed so that the population of faces one encounters most often can be efficiently individuated (Balas, 2012; Valentine, 1991; Valentine et al., 2016). We suggest that Mind-space is similarly experience-dependent, such that the structure of Mind-space reflects the population of minds to which an individual has been exposed. One's developmental experience of different minds would therefore determine the number and type of possible dimensions, and the co-variance between dimensions in Mind-space, in order to enable efficient representation and individuation of the type of minds frequently encountered (Astuti, 2015). Once an individual has constructed their Mind-space then they must learn the mean and variance of each mind they encounter on each of the multiple dimensions and revise the structure of their Mind-space where necessary.

Such an optimization process within Face-space is thought to be responsible for the own-ethnicity advantage to face recognition (Chiroro & Valentine, 1995; Valentine & Endo, 1992), whereby one is better able to individuate faces from one's own ethnic background than those from another ethnic background. It is argued that the number, type, co-variance, and scaling of dimensions are optimized according to the population of faces most commonly experienced (typically from one's own ethnicity), and therefore this space is not optimized to individuate faces drawn from another population (i.e., from a set of other-ethnicity faces), which require a different Face-space structure for optimal individuation. Although experience requiring the individuation of other-ethnicity faces improves this ability, it is interesting to note that this type of experience results in a small decrement in the ability to recognize own-ethnicity faces (Chiroro & Valentine, 1995), presumably as Face-space is no longer perfectly optimized for either population but instead optimized for best performance across the two populations of faces (Valentine et al., 2016).

An analogous process within Mind-space would result in poor models of minds that deviate from the population of minds that one normally encounters. Indeed, Happé and Frith (1996) suggested that children who grow up in abusive or neglectful homes

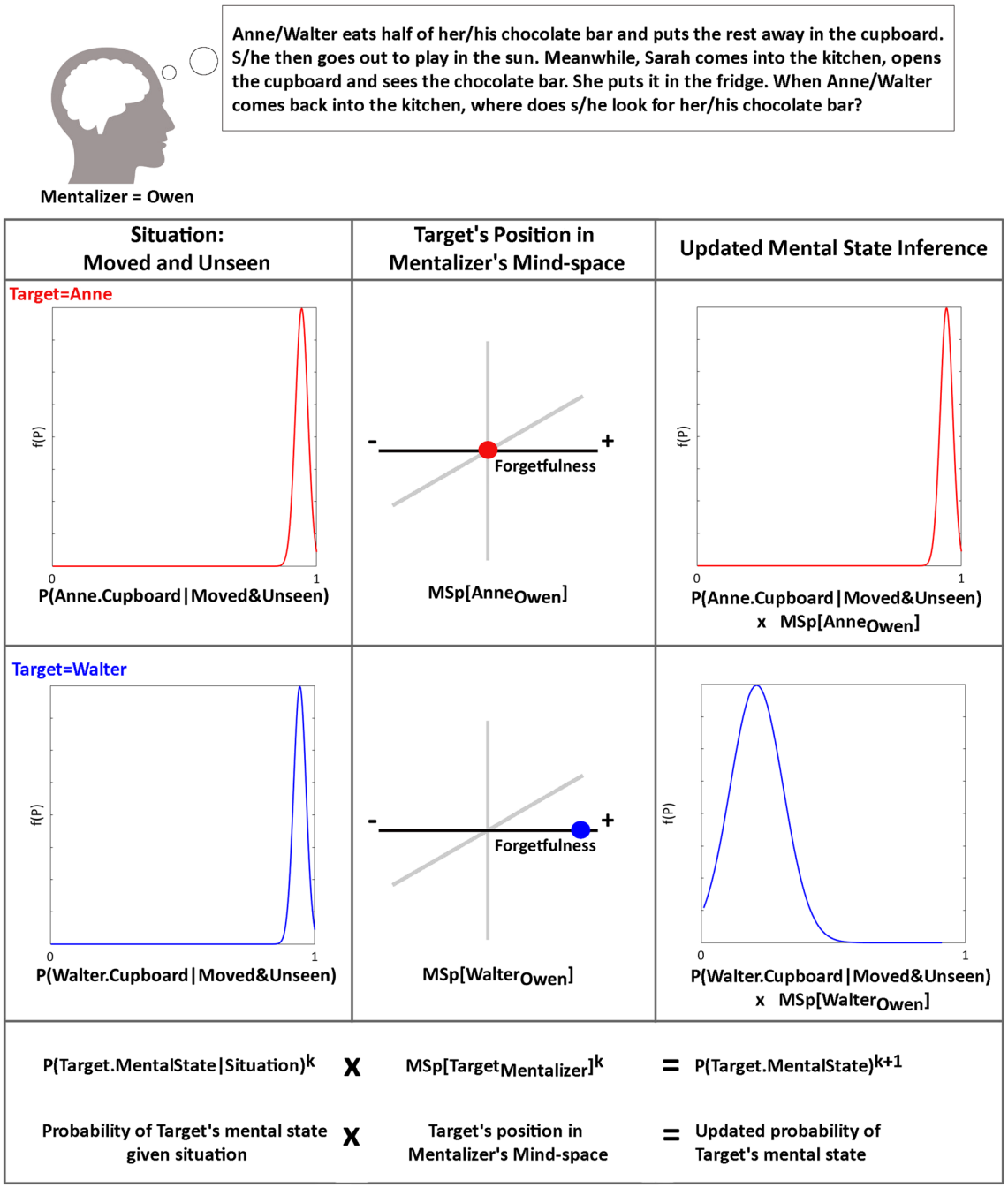

Fig. 3 The relationship between situation, Mind-space, and mental state inference. An example of how the situational factors and location of a target in a mentalizer's Mind-space predict the probability of the mental state content inferred (k = sampling time). Based only on the situational factors, Owen (the mentalizer) predicts that both Anne and Walter are likely to look for their chocolate in the cupboard. Considering their respective positions in Owen's Mind-space on the forgetfulness dimension, Owen revises his prediction for Walter, who is very forgetful and therefore less likely to remember he left the chocolate in the cupboard.

and who are later diagnosed with Conduct Disorder may have developed a model of "nasty" minds, where they overestimate the tendency of others to have minds characterized by aggression, deceitfulness, and a lack of empathy. This model of nasty minds may cause them to be more likely to react with aggression and suspicion when dealing with others, even in the absence of aggression directed towards them. In a similar vein, Frankenhuis and colleagues discuss why those who experienced early life stress such as violence in the home can be faster to identify threat and anger, and better at inferring social dominance and group hierarchy, than those without such developmental experience (Frankenhuis & de Weerth, 2013; Frankenhuis & Del Giudice, 2012; Frankenhuis, Panchanathan, & Nettle, 2016). Less pathologically, optimization of Mind-space for one's own social group

may lead to poor appreciation and understanding of the points of view of those who differ in age, political outlook, culture, or level of education from one's own group, and/or a failure of negotiation when dealing with unfamiliar others.

Inter-group contact has been repeatedly demonstrated to improve the ability of different groups to understand each other's views, reduce stereotyping and increase individuation (Brambilla, Ravenna, & Hewstone, 2012; Bruneau & Saxe, 2012; Harwood, Hewstone, Paolini, & Voci, 2005; Schmid, Ramiah, & Hewstone, 2014), and this may be because such experience allows the modification of Mind-space for efficient representation and individuation of minds dissimilar to those experienced throughout one's developmental history. Indeed, the development and use of stereotypes may reflect poor calibration of Mind-space and a resultant lack of individuation for members of groups other than one's own. If Mind-space works in the same way as Face-space, then the prediction would be that recalibration of Mind-space in response to a distinct population of minds would also result in a small reduction in ability to model the original population of minds, if optimization of Mind-space for both populations of minds results in a sub-optimal space for each independent population (Chiroro & Valentine, 1995; Valentine et al., 2016). A restructuring of Mind-space may serve as a psychological or neurological marker of the reduction in inter-group conflict following inter-group contact.

The experience-dependent nature of Mind-space, and the fact that the accuracy of any particular mental state inference will depend on the quality of the model of a particular mind, means that it becomes less meaningful to talk of an individual or group's 'theory of mind ability' in general terms. A specific individual may be able to infer the contents of a particular target's mental states very well, yet be poor at inferring those of a different target. This can be demonstrated empirically; although typical individuals may exhibit a high degree of accuracy when inferring the mental states of other typical individuals, they are less good at recognizing the emotions (Brewer et al., 2016; Macdonald et al., 1989; Volker, Lopata, Smith, & Thomeer, 2009) and mental states (Edey et al., 2016) of individuals with Autism Spectrum Disorder. To some extent however, a degree of general 'theory of mind ability' (whether good or poor) might be expected due to individual differences in the propensity to model other minds before inferring their mental states, or individual differences in social attention (Chevallier, Molesworth, & Happé, 2012) or social learning (Cook, den Ouden, Heyes, & Cools, 2014), which may impact the speed and quality of learning required to develop Mind-space itself and/or accurately locate an individual target mind within Mind-space. Thus, although the ability to represent minds and the propensity to do so are logically distinct, a greater propensity to represent minds may provide more opportunity for experience-dependent tuning of one's Mind-space, which, given an appropriate learning environment, would increase the accuracy of mind representation and mental state inference.

Some of the strongest evidence for the experience-dependent and dimensional aspects of Face-space comes from adaptation effects. Face adaptation occurs when exposure to faces at extreme ends of a dimension, such as attractiveness (Rhodes et al., 2003), gender (Webster et al., 2004), or contractedness (Jeffery & Rhodes, 2011), shifts the mean of that dimension such that stimuli originally perceived as neutral subsequently appear further from the adapting face. For example, prolonged exposure to a very wide face will mean that other faces are perceived as narrower than before the exposure to the wide face. There is some indirect evidence that adaptation may also occur in Mind-space; Xiang and colleagues (2013) demonstrated that exposure to generous or unfair offers in an Ultimatum Game affected subsequent rejection rates and mood ratings for fair, neutral and unfair offers. Directly testing for adaptation effects in Mind-space would provide a strong test of whether minds are represented along dimensions (Heleven & Van Overwalle, 2018; Ma et al., 2014), rather than categories, and whether experience affects the structure of Mind-space.

Typical and atypical development of Mind-space

We have suggested that the development of Mind-space is experience-dependent. Typical developmental effects in the ability to represent minds and accurately infer the content of mental states may reflect the formation of a higher-dimensional Mind-space, more appropriate weighting of dimensions, and/or an increasing ability or propensity to locate individuals within Mind-space. Indeed, considering atypical development of Mind-space provides for the establishment of further sources of individual differences in mental state inference. Over development, one must learn the relative importance of different dimensions of Mind-space in determining mental states in particular contexts, and how variance in these dimensions predicts variance in mental states. Atypical experience may lead to atypical mental state inferences even when the target is located correctly in a typical Mind-space. For example, if a child grew up in a family with a depressed parent who exhibited atypical depression-related mental states (i.e., atypical within the population of depressed individuals), then they may learn an atypical model of how position on the depression dimension of Mind-space predicts the likelihood of specific types of mental state. If they subsequently encounter a second depressed individual, who they correctly locate on the depression dimension in their Mind-space, then if they apply this atypical 'Mind-space to mental state' model to the second depressed individual they will make an incorrect inference regarding their mental state. Additionally, if mind representation is culturally acquired, then the Mind-space framework is sufficiently flexible to account for cultural differences in how minds are represented. Theories of minds change across cultures

(Lillard, 1998; Perez-Zapata, Slaughter, & Henry, 2016; Shahaeian et al., 2011), and perhaps across historical time, and therefore any psychological model of how minds are represented needs to account for varying concepts of mind.

Mind-space provides a framework for investigating the development of advanced social skills; for example, the ability to quickly extract diagnostic information to locate someone within Mind-space. Conversely, the Mind-space framework may shed light on the social impairments which are a transdiagnostic trait of many psychiatric and neurodevelopmental disorders, including autism, depression, eating disorders, and personality disorders (Happé, 1994; Preißler, Dziobek, Ritter, Heekeren, & Roepke, 2010; Russell, Schmidt, Doherty, Young, & Tchanturia, 2009; Wang, Wang, Chen, Zhu, & Wang, 2008). Under this framework social impairments may reflect: (1) an atypical representation of Mind-space (for example, the paranoia observed in schizophrenia (Drake et al., 2004) could reflect a misaligned, over-weighted, or otherwise atypical dimension representing others' hostility); (2) a decreased propensity to model other minds; or (3) a fundamentally altered learning system that results in decreased generalization of learning (e.g., Plaisted, 2001), or a reduced influence of priors (Pellicano & Burr, 2012), which impacts the updating of Mind-space from experience. For example, it has been claimed that individuals with autism show insufficient generalization of their learning (Plaisted, 2001). As a consequence, autistic individuals may be too specific in their mental models, failing to generalize across instances to develop population-based representations of Mind-space. Conversely, some of the social difficulties encountered by individuals diagnosed with psychiatric conditions may be caused by a failure of typical individuals to be able to develop an accurate model of atypical minds (Brewer et al., 2016; Edey et al., 2016; Sasson et al., 2017).

Concluding remarks

In this article we sought to address an impasse in the theory of mind literature, specifically the inability of current frameworks to characterize individual differences in theory of mind ability, and to introduce a framework within which all aspects of minds can be represented. We have suggested that the adoption of a Mind-space framework where minds are represented within a multidimensional space – similar to that which has been so successful in providing a unifying theoretical framework for the study of faces – would achieve both aims. Mind-space represents a psychological model of a representational structure involved in the representation of minds, which may also explain variance in the accuracy of mental state inference. It considers how individuals build models of other minds, and suggests that there may be substantial variance in the accuracy of mental state inference within an individual based on the quality of their representation of the target mind. Future work can determine whether analogous effects to those in the face processing literature explained by Face-space can be observed for mind representation by adopting the Mind-space framework. Findings equivalent to the own-ethnicity bias and perceptual adaptation seen in faces would explain much about how inter-group conflict may be generated, maintained, and reduced. We hope that this introductory sketch of Mind-space is a first step towards an understanding of individual differences in the representation of whole cognitive systems, where minds are recognized as complex multidimensional stimuli. It should be noted, however, that even if minds are not represented in a multidimensional space, the ability and propensity to represent another's mind is still likely to be an important source of individual differences in the accuracy of mental state inference.

Acknowledgements The authors are very grateful to Dr Michel-Pierre Coll for helpful discussion of earlier drafts of this manuscript. A brief version of this paper was presented at "*A Penny for Your Thoughts: A Workshop on Social Cognition,*" University of London, 25 September 2017 (https://www.sas.ac.uk/videos-and-podcasts/philosophy/penny-your-thoughts-workshop-social-cognition-1). This work was supported by an Economic and Social Research Council studentship [Ref: 1413340] awarded to J.R. Conway.

References

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

Astuti, R. (2015). Implicit and explicit theory of mind. *Anthropology of this Century, 13*. Retrieved from http://aotcpress.com/articles/implicit-explicit-theory-mind/

Baker, C. L., Jara-Ettinger, J., Saxe, R., & Tenenbaum, J. B. (2017). Rational quantitative attribution of beliefs, desires and percepts in human mentalizing. *Nature Human Behaviour*, *1*(4), 0064. https://doi.org/10.1038/s41562-017-0064

Balas, B. (2012). Bayesian face recognition and perceptual narrowing in face-space. *Developmental Science*, *15*(4), 579–588. https://doi.org/10.1111/j.1467-7687.2012.01154.x

Baron-Cohen, S., Leslie, A. M., & Frith, U. (1985). Does the autistic child have a "theory of mind"? *Cognition*, *21*(1), 37–46. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/9775957

Bartsch, K., & Estes, D. (1996). Individual differences in children's developing theory of mind and implications for metacognition. *Learning and Individual Differences*, *8*(4), 281–304. https://doi.org/10.1016/S1041-6080(96)90020-5

Bird, G. (2017). '*A Penny for Your Thoughts: A Workshop on Social Cognition*', University of London, 25th Sept 2017. Retrieved from https://www.sas.ac.uk/videos-and-podcasts/philosophy/penny-your-thoughts-workshop-social-cognition-1

Blokpoel, M., van Kesteren, M., Stolk, A., Haselager, P., Toni, I., & van Rooij, I. (2012). Recipient design in human communication: simple heuristics or perspective taking? *Frontiers in Human Neuroscience*, *6*(September), 1–13. https://doi.org/10.3389/fnhum.2012.00253

Brambilla, M., Ravenna, M., & Hewstone, M. (2012). Changing stereotype content through mental imagery: Imagining intergroup contact promotes stereotype change. *Group Processes & Intergroup Relations*, *15*(3), 305–315. https://doi.org/10.1177/1368430211427574

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

Butterfill, S. A., & Apperly, I. A. (2013). How to construct a minimal theory of mind. *Mind & Language*, *28*(5), 606–637. https://doi.org/10.1111/mila.12036

Camerer, C. F., Ho, T. H., & Chong, J. K. (2004). A cognitive hierarchy model of games. *The Quarterly Journal of Economics*, *119*(3), 861-898.

Cai, Z. G., Gilbert, R. A., Davis, M. H., Gaskell, M. G., Farrar, L., Adler, S., & Rodd, J. M. (2017). Accent modulates access to word meaning: Evidence for a speaker-model account of spoken word recognition. *Cognitive Psychology*, *98*, 73–101. https://doi.org/10.1016/j.cogpsych.2017.08.003

Carlson, S. M., & Moses, L. J. (2001). Individual Differences in Inhibitory Control and Children's Theory of Mind. *Child Development, 72*(4), 1032–1053.

Carruthers, P., & Smith, P. K. (Eds.). (1996). Theories of theories of mind. Cambridge, UK: Cambridge University Press.

Chevallier, C., Molesworth, C., & Happé, F. (2012). Diminished social motivation negatively impacts reputation management: Autism spectrum disorders as a case in point. *PLoS ONE*, *7*(1), e31107. https://doi.org/10.1371/journal.pone.0031107

Chiroro, P., & Valentine, T. (1995). An investigation of the contact hypothesis of the own-race bias in face recognition. *The Quarterly Journal of Experimental Psychology*, *48*(4), 879–894. https://doi.org/10.1080/14640749508401421

Conway, J. R., & Bird, G. (2018). Conceptualizing degrees of theory of mind. *Proceedings of the National Academy of Sciences*, *115*(7), 201722396. https://doi.org/10.1073/pnas.1722396115

Cook, J. L., den Ouden, H. E. M., Heyes, C. M., & Cools, R. (2014). The social dominance paradox. *Current Biology*, *24*(23), 2812–2816. https://doi.org/10.1016/j.cub.2014.10.014

Coricelli, G., & Nagel, R. (2009). Neural correlates of depth of strategic reasoning in medial prefrontal cortex. *Proceedings of the National Academy of Sciences*, *106*(23), 9163-9168. https://doi.org/10.1073/pnas.0807721106

de Heering, A., Rossion, B., & Maurer, D. (2012). Developmental changes in face recognition during childhood: Evidence from upright and inverted faces. *Cognitive Development*, *27*(1), 17–27. https://doi.org/10.1016/j.cogdev.2011.07.001

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

Drake, R., Pickles, A., Bentall, R. P., Kinderman, P., Haddock, G., Tarrier, N., & Lewis, S. (2004). The evolution of insight, paranoia and depression during early schizophrenia. *Psychological Medicine*, *34*(2), 285–292.

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

Fiske, S. T., Cuddy, A. J. C., & Glick, P. (2007). Universal dimensions of social cognition: warmth and competence. *Trends in Cognitive Sciences*, *11*(2), 77–83. https://doi.org/10.1016/j.tics.2006.11.005

Frankenhuis, W. E., & de Weerth, C. (2013). Does early-life exposure to stress shape or impair cognition? *Current Directions in Psychological Science*, *22*(5), 407–412. https://doi.org/10.1177/0963721413484324

Frankenhuis, W. E., & Del Giudice, M. (2012). When do adaptive developmental mechanisms yield maladaptive outcomes? *Developmental Psychology*, *48*(3), 628–642. https://doi.org/10.1037/a0025629

Frankenhuis, W. E., Panchanathan, K., & Nettle, D. (2016). Cognition in harsh and unpredictable environments. *Current Opinion in Psychology*, *7*, 76–80. https://doi.org/10.1016/j.copsyc.2015.08.011

Frith, U. (2003). *Autism: Explaining the Enigma* (2nd ed.). Wiley-Blackwell.

Gray, H. M., Gray, K., & Wegner, D. M. (2007). Dimensions of mind perception. *Science (New York, N.Y.)*, *315*(5812), 619. https://doi.org/10.1126/science.1134475

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), 129–54. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/8040158

Happé, F. G., Cook, J. L., & Bird, G. (2017). The structure of social cognition : In(ter) dependence of sociocognitive processes. *Annu. Rev. Psychol*, *68*(September 2016), 1–25. https://doi.org/10.1146/annurev-psych-010416-044046

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.

Harwood, J., Hewstone, M., Paolini, S., & Voci, A. (2005). Grandparent-grandchild contact and attitudes toward older adults: moderator and mediator effects. *Personality & Social Psychology Bulletin*, *31*(3), 393–406. https://doi.org/10.1177/0146167204271577

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. https://doi.org/10.1093/cercor/bht042

Heleven, E., & Van Overwalle, F. (2015). The person within: Memory codes for persons and traits using fMRI repetition suppression. *Social Cognitive and Affective Neuroscience*, *11*(1), 159–171. https://doi.org/10.1093/scan/nsv100

Heleven, E., & Van Overwalle, F. (2018). The neural basis of representing others' inner states. *Current Opinion in Psychology*, *23*, 98–103. https://doi.org/10.1016/j.copsyc.2018.02.003

Heyes, C. M., & Frith, C. D. (2014). The cultural evolution of mind reading. *Science*, *344*(6190), 1243091. https://doi.org/10.1126/science.1243091

Hills, P. J., Holland, A. M., & Lewis, M. B. (2010). Aftereffects for face attributes with different natural variability: Children are more adaptable than adolescents. *Cognitive Development*, *25*(3), 278–289. https://doi.org/10.1016/j.cogdev.2010.01.002

Hughes, C. (1998). Executive function in preschoolers: Links with theory of mind and verbal ability. *British Journal of Developmental Psychology*, *16*(2), 233–253.

Jacob, P., & Jeannerod, M. (2005). The motor theory of social cognition: A critique. *Trends in Cognitive Sciences, 9*(1), 21-25. https://doi.org/10.1016/j.tics.2004.11.003

Jeffery, L., & Rhodes, G. (2011). Insights into the development of face recognition mechanisms revealed by face aftereffects. *British Journal of Psychology*, *102*(4), 799–815. https://doi.org/10.1111/j.2044-8295.2011.02066.x

Jiang, F., Blanz, V., & O'Toole, A. J. (2006). Probing the visual representation of faces with adaptation: A view from the other side of the mean. *Psychological Science*, *17*(6), 493–500. https://doi.org/10.1111/j.1467-9280.2006.01734.x

Jiang, F., Blanz, V., & O'Toole, A. J. (2009). Three-dimensional information in face representations revealed by identity aftereffects. *Psychological Science*, *20*(3), 318–325. https://doi.org/10.1111/j.1467-9280.2009.02285.x

Kemper, S., & Harden, T. (1999). Experimentally disentangling what's beneficial about elderspeak from what's not. *Psychology and Aging*, *14*(4), 656–668.

Leopold, D. A., O'Toole, A. J., Vetter, T., & Blanz, V. (2001). Prototype-referenced shape encoding revealed by high-level aftereffects. *Nature Neuroscience*, *4*(1), 89–94. https://doi.org/10.1038/82947

Leslie, A. M. (1987). Pretense and representation: The origins of theory of mind. *Psychological Review*, *94*(4), 412–426. https://doi.org/10.1037/0033-295X.94.4.412

Lillard, A. (1998). Ethnopsychologies: Cultural variations in theories of mind,. *Psychological Bulletin*, *123*(1), 3–32. https://doi.org/10.1037/0033-2909.123.1.3

Ma, N., Baetens, K., Vandekerckhove, M., Kestemont, J., Fias, W., & Van Overwalle, F. (2014). Traits are represented in the medial prefrontal cortex: An fMRI adaptation study. *Social Cognitive and Affective Neuroscience*, *9*(8), 1185–1192. https://doi.org/10.1093/scan/nst098

MacDonald, H., Rutter, M., Howlin, P., Rios, P., Le Conteur, A., Evered, C., & Folstein, S. (1989). Recognition and expression of emotional cues by autistic and normal adults. *Journal of Child Psychology and Psychiatry and Allied Disciplines*, *30*(6), 865–877. https://doi.org/10.1111/j.1469-7610.1989.tb00288.x

McCrae, R. R. (1989). Why I advocate the five-factor model: Joint analyses of the NEO-PI with other instruments. In D. M. Buss & N. Cantor (Eds.), Personality Psychology: Recent trends and emerging directions (pp. 237-245). New York: Springer-Verlag.

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

Newman-Norlund, S. E., Noordzij, M. L., Newman-Norlund, R. D., Volman, I. A. C., Ruiter, J. P. de, Hagoort, P., & Toni, I. (2009). Recipient design in tacit communication. *Cognition*, *111*(1), 46–54. https://doi.org/10.1016/j.cognition.2008.12.004

Oakley, B. F. M., Brewer, R., Bird, G., & Catmur, C. (2016). 'Theory of Mind' is not theory of emotion: A cautionary note on the reading the mind in the eyes test. *Journal of Abnormal Psychology*, *125*(6), 818–23. https://doi.org/10.1037/abn0000182

Oosterhof, N. N., & Todorov, A. (2008). The functional basis of face evaluation. *Proceedings of the National Academy of Sciences of the United States of America*, *105*(32), 11087–92. https://doi.org/10.1073/pnas.0805664105

Pellicano, E., & Burr, D. (2012). When the world becomes "too real": A Bayesian explanation of autistic perception. *Trends in Cognitive Sciences*, *16*(10), 504–510. https://doi.org/10.1016/j.tics.2012.08.009

Perez-Zapata, D., Slaughter, V., & Henry, J. D. (2016). Cultural effects on mindreading. *Cognition*, *146*, 410–414. https://doi.org/10.1016/j.cognition.2015.10.018

Plaisted, K. C. (2001). Reduced generalization in autism: An alternative to Weak Central Coherence. In J. Burack, T. Charman, N. Yirmiya, & P. Zelazo (Eds.), The development of autism: Perspectives from theory and research (pp. 149–169). Lawrence Erlbaum Associates Publishers.

Preißler, S., Dziobek, I., Ritter, K., Heekeren, H. R., & Roepke, S. (2010). Social cognition in borderline personality disorder : evidence for disturbed recognition of the emotions , thoughts, and intentions of others. *Frontiers in Behavioral Neuroscience*, *4*(December), 1–8. https://doi.org/10.3389/fnbeh.2010.00182

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

Rhodes, G., Jeffery, L., Watson, T. L., Clifford, C. W. G., & Nakayama, K. (2003). Fitting the mind to the world: Face adaptation and attractiveness aftereffects. *Psychological Science*, *14*(6), 558–567.

Russell, T. A., Schmidt, U., Doherty, L., Young, V., & Tchanturia, K. (2009). Aspects of social cognition in anorexia nervosa : Affective and cognitive theory of mind. *Psychiatry Research*, *168*(3), 181–185. https://doi.org/10.1016/j.psychres.2008.10.028

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.

Samson, D., Apperly, I. A., Braithwaite, J. J., Andrews, B. J., & Bodley Scott, S. E. (2010). Seeing it their way: evidence for rapid and involuntary computation of what other people see. *Journal of Experimental Psychology. Human Perception and Performance*, *36*(5), 1255–66. https://doi.org/10.1037/a0018729

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

Schaafsma, S. M., Pfaff, D. W., Spunt, R. P., & Adolphs, R. (2015). Deconstructing and reconstructing theory of mind. *Trends in Cognitive Sciences*, *19*(2), 65–72. https://doi.org/10.1016/j.tics.2014.11.007

Schmid, K., Ramiah, A. Al, & Hewstone, M. (2014). Neighborhood ethnic diversity and trust: The role of intergroup contact and perceived threat. *Psychological Science*, *25*(3), 665–674. https://doi.org/10.1177/0956797613508956

Shahaeian, A., Nielsen, M., Peterson, C. C., Aboutalebi, M., & Slaughter, V. (2014a). Knowledge and belief understanding among Iranian and Australian preschool children. *Journal of Cross-Cultural Psychology*, *45*, 1643–1654. https://doi.org/10.1177/0022022114548484

Shahaeian, A., Nielsen, M., Peterson, C. C., & Slaughter, V. (2014b). Cultural and family influences on children's theory of mind development: A comparison of Australian and Iranian school-age children. *Journal of Cross-Cultural Psychology*, *45*(4), 555–568. https://doi.org/10.1177/0022022113513921

Shahaeian, A., Peterson, C. C., Slaughter, V., & Wellman, H. M. (2011). Culture and the sequence of steps in theory of mind development. *Developmental Psychology*, *47*(5), 1239–1247. https://doi.org/10.1037/a0023899

Slaughter, V., & Perez-Zapata, D. (2014). Cultural variations in the development of mind reading. *Child Development Perspectives*, *8*(4), 237–241. https://doi.org/10.1111/cdep.12091

Spunt, R. P., & Adolphs, R. (2015). Folk explanations of behavior: a specialized use of a domain-general mechanism. *Psychological Science*, *26*(6), 724–736. https://doi.org/10.1177/0956797615569002

Stolk, A., Hunnius, S., Bekkering, H., & Toni, I. (2013). Early social experience predicts referential communicative adjustments in five-year-old children. *PLoS ONE*, *8*(8), e72667. https://doi.org/10.1371/journal.pone.0072667

Stolk, A., Noordzij, M. L., Verhagen, L., Volman, I., Schoffelen, J.-M., Oostenveld, R., … Toni, I. (2014a). Cerebral coherence between communicators marks the emergence of meaning. *Proceedings of the National Academy of Sciences*, *111*(51), 18183–18188. https://doi.org/10.1073/pnas.1414886111

Stolk, A., Noordzij, M. L., Volman, I., Verhagen, L., Overeem, S., van Elswijk, G., … Toni, I. (2014b). Understanding communicative actions: A repetitive TMS study. *Cortex*, *51*(1), 25–34. https://doi.org/10.1016/j.cortex.2013.10.005

Tamir, D. I., & Thornton, M. A. (2018). Modeling the predictive social mind. *Trends in Cognitive Sciences*, *22*(3), 201–212. https://doi.org/10.1016/j.tics.2017.12.005

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. https://doi.org/10.1073/pnas.1511905112

Thompson, P., & Burr, D. (2009). Visual aftereffects. *Current Biology*, *19*(1), 11–14. https://doi.org/10.1016/j.cub.2008.10.014

Todorov, A., Olivola, C. Y., Dotsch, R., & Mende-Siedlecki, P. (2015). Social attributions from faces: Determinants, consequences, accuracy, and functional significance. *Annu. Rev. Psychol*, *66*, 519–545. https://doi.org/10.1146/annurev-psych-113011-143831

Todorov, A., Said, C. P., Engell, A. D., & Oosterhof, N. N. (2008). Understanding evaluation of faces on social dimensions. *Trends in Cognitive Sciences*, *12*(12), 455–460. https://doi.org/10.1016/j.tics.2008.10.001

Valentine, T. (1991). A unified account of the effects of distinctiveness, inversion, and race in face recognition. *The Quarterly Journal of Experimental Psychology*, *43:2*(July), 161–204. https://doi.org/10.1080/14640749108400966

Valentine, T., & Endo, M. (1992). Towards an exemplar model of face processing: The effects of race and distinctiveness. *Quarterly Journal of Experimental Psychology*, *44*(4), 671–703. https://doi.org/10.1080/14640749208401305

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. https://doi.org/10.1080/17470218.2014.990392

Volker, M. A., Lopata, C., Smith, D. A., & Thomeer, M. L. (2009). Facial encoding of children with high-functioning autism spectrum disorders. *Focus on Autism and Other Developmental Disabilities*, *24*(4), 195–204. https://doi.org/10.1177/1088357609347325

Wang, Y., Wang, Y., Chen, S., Zhu, C., & Wang, K. (2008). Theory of mind disability in major depression with or without psychotic symptoms : A componential view. *Psychiatry Research*, *161*(2), 153–161. https://doi.org/10.1016/j.psychres.2007.07.018

Webster, M. A., Kaping, D., Mizokami, Y., & Duhamel, P. (2004). Adaptation to natural facial categories. *Nature*, *428*(April), 557–561. https://doi.org/10.1038/nature02361.1.

Wellman, H. M., Fang, F., Liu, D., Zhu, L., & Liu, G. (2006). Scaling of theory-of-mind understandings in Chinese children. *Psychological Science*, *17*(12), 1075–1081. https://doi.org/10.1111/j.1467-9280.2006.01830.x

Wellman, H. M., Fang, F., & Peterson, C. C. (2011). Sequential progressions in a theory of mind scale: Longitudinal Perspectives. *Child Development, 82*(3), 780–792. https://doi.org/10.1111/j.1467-8624.2011.01583.x.

Wellman, H. M., & Liu, D. (2004). Scaling of theory-of-mind tasks. *Child Development*, *75*(2), 523–541. https://doi.org/10.1111/j.1467-8624.2004.00691.x

Williams, K., Kemper, S., & Hummert, M. L. (1995). Practice concepts improving nursing home communication : An intervention to reduce elderspeak. *The Gerontologist*, *43*(2), 242–247. https://doi.org/10.1093/geront/43.2.242

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.

Xiang, T., Lohrenz, T., & Montague, P. R. (2013). Computational substrates of norms and their violations during social exchange. *The Journal of Neuroscience*, *33*(3), 1099–1108. https://doi.org/10.1523/JNEUROSCI.1642-12.2013

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

This chapter is presented as an article that is currently under review and has incorporated revisions invited by the editor:

Conway, J.R., Coll, M-P., Cuve, H.C., Koletsi, S., Bronitt, N., Catmur, C., & Bird, G. (*under review*). Understanding How Minds Vary Relates to Skill in Inferring Mental States, Personality and Intelligence. *Journal of Experimental Psychology: General*.

Corresponding author:

J.R. Conway, Social, Genetic and Developmental Psychiatry Centre, Institute of Psychiatry, Psychology and Neuroscience, King's College London, London SE5 8AF, United Kingdom.

jane_rebecca.conway@kcl.ac.uk

3.1 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.

3.2 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 mediate 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 prefer (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 multidimensional space where dimensions represent ways in which faces can be discriminated. Once formed, individual faces are thought to be represented as locations within this multidimensional 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 multidimensional 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 (see Figure 3.1). 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 3.1 for a full example.)

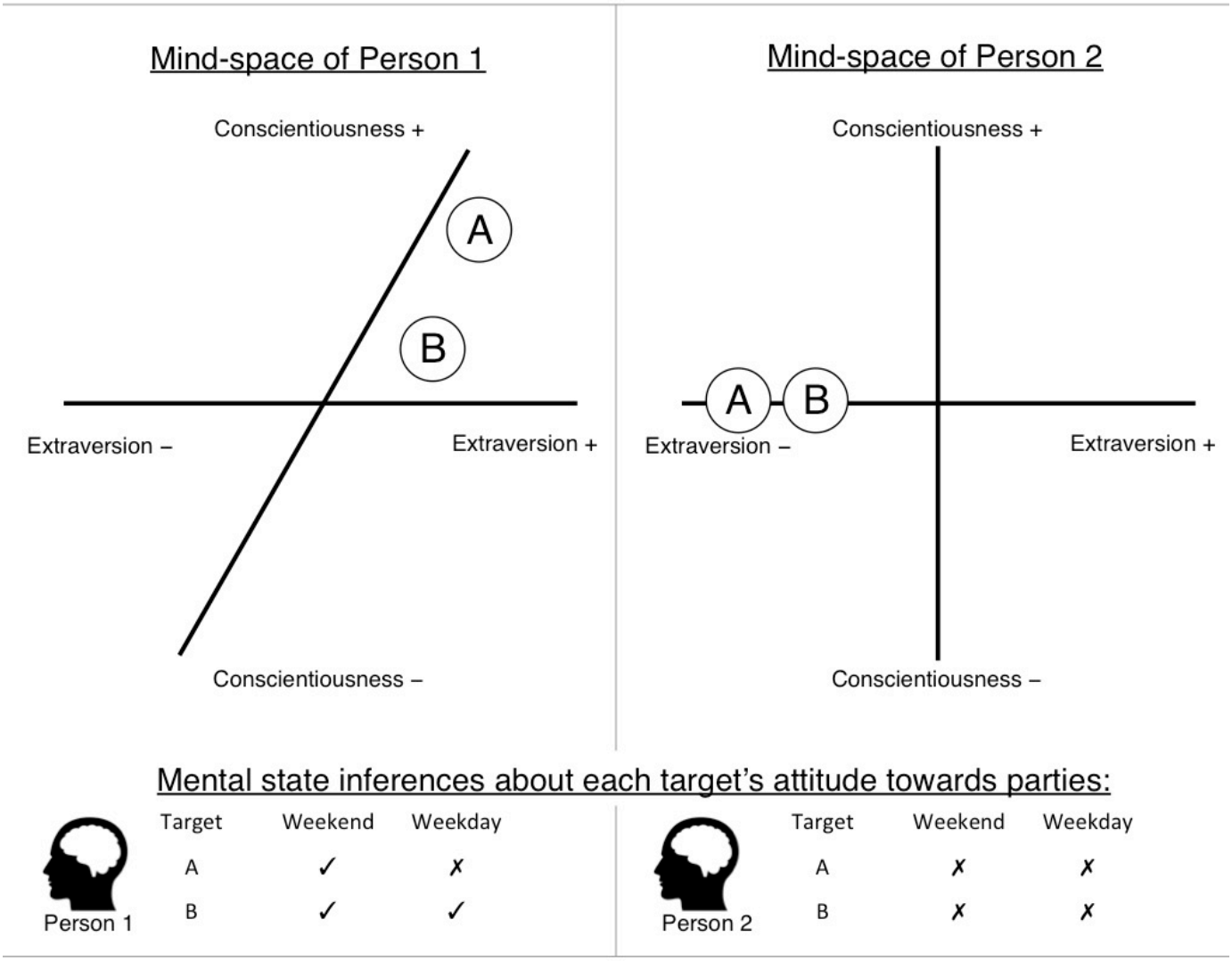

Figure 3.1 Schematic illustration of how the Mind-space framework can be used to explain individual differences in Theory of Mind (ToM).
The Mind-space framework suggests that individual differences in ToM are due to: (1) The accuracy of the representation of the dimensions within which minds vary and 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 multidimensional 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 taken 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 multidimensional 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 multidimensional 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.

3.3 Experiment 1

3.3.1 Method

3.3.1.1 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.

3.3.1.2 Measures

3.3.1.2.1 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.

3.3.1.2.2 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).

3.3.1.3 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].

3.3.1.4 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.

3.3.2 Results

Descriptive statistics for all variables are presented in Table 3.1. 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 3.2, 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 3.2 (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 3.2, 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.

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

Table 3.1 Descriptive Statistics for Experiment 1
PPT = Personality Pairs Task. MS = Mental State. Qs = Questions.

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

Table 3.2 Experiment 1 Regression Analyses: Predictors of Performance on the Personality Pairs Task
Qs = Questions. MS = Mental State. * $p < .05$. ** $p < .01$.

3.3.3 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.

3.4 Experiment 2

3.4.1 Method

3.4.1.1 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.

3.4.1.2 Measures

3.4.1.2.1 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.

3.4.1.2.2 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.

3.4.1.2.3 Personality Pairs Task

As described in Experiment 1 (3.3.1.2.1)

3.4.1.2.4 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.

3.4.1.2.5 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].

3.4.1.2.6 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.

3.4.2 Results

Descriptive statistics for all variables are presented in Table 3.3. 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 (Figure 3.4 and Table 3.7).

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

Table 3.3 Descriptive Statistics for Experiment 2
PPT = Personality Pairs Task. SRH = Self-report HEXACO. ORH = Observer-report HEXACO. IQ = Intelligence.

As shown in Table 3.4, 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 3.5, 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.

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*

Table 3.4 Experiment 2: Regression Analyses.
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. * $p < .05$. ** $p < .001$.

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

Table 3.5 Experiment 2: Regression Analyses.
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. ** $p < .001$.

3.4.3 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.

3.5 Experiment 3

3.5.1 Method

3.5.1.1 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.

3.5.1.2 Measures

3.5.1.2.1 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.

3.5.1.2.2 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.

3.5.1.2.3 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.

3.5.1.2.4 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.

3.5.1.2.5 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.

3.5.1.2.6 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.

3.5.1.2.7 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.

3.5.1.2.8 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.

3.5.2 Results

Descriptive statistics for all variables are presented in Table 3.6. 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 3.8). 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 3.9). The effects of paranoia and dishonesty on the probability of mental state inferences are shown in Figure 3.2.

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 3.9. 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 Table 3.10 and Table 3.11. 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.

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

Table 3.6 Descriptive Statistics for Experiment 3.
Higher values on Mental State Probability indicate a higher probability of the false belief location. SD = Standard Deviation.

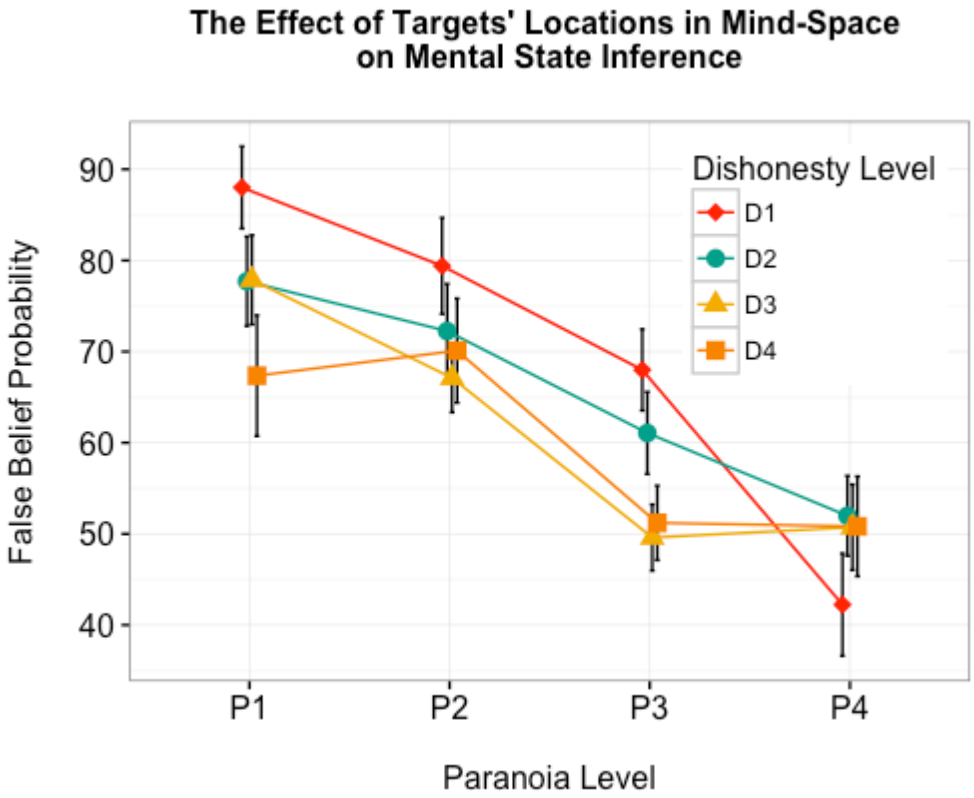

Figure 3.2 The effect of targets' locations in Mind-space on the probability of the mental state inferred.
Note that higher values on the 'False Belief Probability' axis indicate higher probabilities of searching in the 'false belief' location, that is, where the Sally character left her object. Error bars show within-subject 95% confidence intervals around the means (Morey, 2008).

3.5.3 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 protagonist's 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 participant's reasoning. If the participant attributes paranoia/dishonesty 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 multidimensional 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 multidimensional 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.

3.6 Experiment 4

3.6.1 Methods

3.6.1.1 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.

3.6.1.2 Measures

3.6.1.2.1 Mental State Stories

The same 32 vignettes used in Section 3.5.1.2.1 were also used in this experiment.

3.6.1.2.2 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 3.5). 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
- trusting

which are negatively correlated with paranoia, and:

- superstitious
- pessimistic
- cautious

which are positively correlated with paranoia.

3.6.1.2.3 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*
- *Very Much*.

At Parnoia 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.

3.6.1.2.4 Mental State Inference

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

3.6.1.2.5 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?"* (Table 3.13). 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.

3.6.1.2.6 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:

- participants locate minds within Mind-space based on information they are given about particular source traits;
- they use that information to locate those minds on dimensions they believe to be correlated with the source traits;
- and 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.

3.6.2 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 3.12. As predicted, the model results show a significant effect of MPRP on the False Belief Probability attribution (β = - 8.85, *95% CI* [-10.65, -7.03], $p < .001$, see Figure 3.3 and Table 3.14). 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 ($\chi^2_{(1)} = 53.50$, $p < .001$, see Table 3.15).

Examination of the AIC and BIC values also showed that the MPRP model outperformed the Sally source traits model (ΔAIC = 32, ΔBIC = 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.

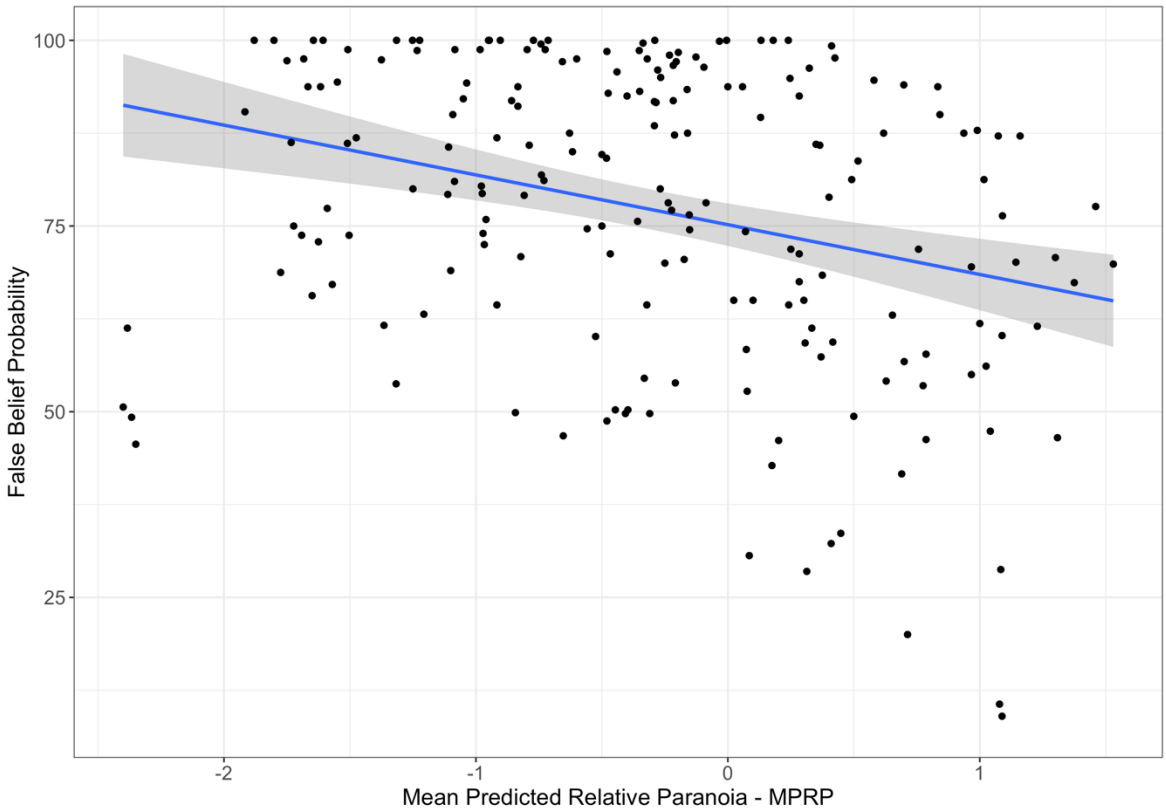

Figure 3.3 Effect of Mean Predicted Relative Paranoia (MPRP) score on 'False Belief' Attribution.
Shaded area represents the 95% confidence interval. MPRP is calculated on the basis of the Sally character's scores on various traits and the degree to which each participant believes those traits to be associated with paranoia.

3.6.3 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).

3.7 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 3.1). 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.

3.8 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.

3.9 Acknowledgments

This work was supported by an Economic and Social Research Council studentship [Ref: 1413340] awarded to J.R. Conway.

3.10 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.

3.11 Supplemental Materials

3.11.1 Supplemental Methods for Experiment 2

Figure 3.4 presents density plots for the distribution of personality traits within the sample tested in Experiment 2. As can be seen, scores on each of the HEXACO scales were similarly distributed.

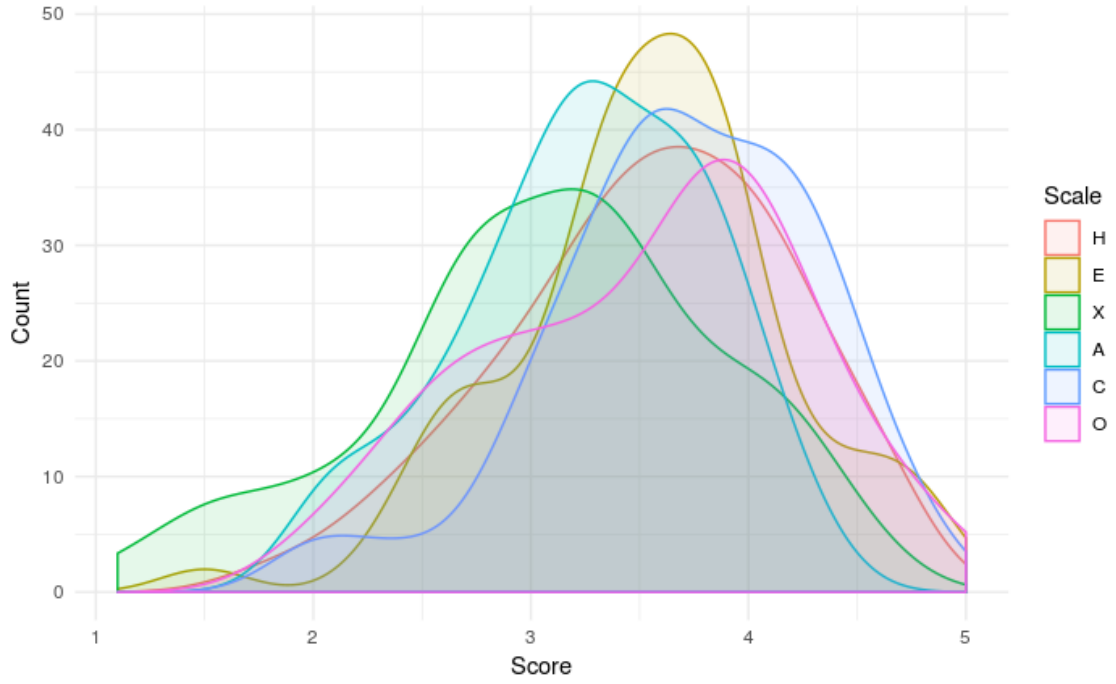

Figure 3.4 Distribution of scores on each scale of the HEXACO.

To assess the possibility that personality traits themselves, and not similarity between the participants and the targets, predicted the accuracy of estimates of personality traits, we performed additional analyses including the 6 scales of the HEXACO as fixed effects in the linear models assessing the relationship between Similarity of the HEXACO scores between participants and targets and accuracy of the estimation of the self-reported (SRH) and other-reported (ORH) targets' HEXACO scores (original models 2.D and 2.E). As shown in the supplementary Table 3.7 showing the results for Model 2.G.1 and 2.H.1, none of the participants' HEXACO traits predicted the accuracy of estimates of personality traits after correcting for

multiple comparisons. Formal model comparisons with and without the Similarity predictor revealed that Similarity significantly increased model fit for both Model 2.G and 2.H after taking into account personality traits.

Predictor	Model 2.G.1 DV: SRHds *B* (95% CI)	Model 2.G.2 DV: SRHds *B* (95% CI)	Model 2.H.1 DV: ORHds *B* (95% CI)	Model 2.H.2 DV: ORHds *B* (95% CI)
Similarity		0.173** (0.153, 0.194)		0.053** (0.033, 0.073)
H	-0.010 (-0.053, 0.032)	-0.012 (-0.054, 0.031)	-0.004 (-0.051, 0.044)	-0.004 (-0.052, 0.044)
E	0.049 (0.003, 0.095)	0.045 (-0.002, 0.091)	0.051 (-0.0004, 0.103)	0.050 (-0.002, 0.102)
X	0.022 (-0.015, 0.059)	0.030 (-0.008, 0.068)	0.024 (-0.019, 0.066)	0.026 (-0.016, 0.068)
A	0.050 (-0.003, 0.104)	0.055 (0.001, 0.109)	0.057 (-0.003, 0.118)	0.069 (0.014, 0.125)
C	0.060 (0.011, 0.109)	0.061 (0.012, 0.111)	0.069 (0.014, 0.124)	-0.036 (0.047, 0.092)
O	-0.023 (-0.062, 0.017)	-0.016 (-0.056, 0.024)	-0.038 (-0.083, 0.007)	-0.036 (-0.081, 0.009)
Δ -2LL		131.00		13.42
$X^2(1)$		261.99 **		26.83**

Table 3.7 Additional multiple linear mixed-models.
These models assess the relationship between the HEXACO personality traits and the self (SRH) and other (ORH) reported differences scores in Experiment 2. The second step of each model tests the effect on model fit of the addition of the Similarity predictor. **$p < 0.001$ after Bonferroni correction; Note: All models included Participant, Target and Trait random effects.

3.11.2 Supplemental Materials for Experiment 3

	Condition	
Levels	Paranoia	Dishonesty
1:2	M_{diff} = 5.52 (2.01), CI [0.03, 11.01], F (1, 62) = 7.51, p = .008, η_p^2 = .11.	M_{diff} = 3.66 (1.39), CI [- 0.13, 7.44], F (1, 62) = 6.94, p = .011, η_p^2 = .10.
1:3	M_{diff} = 20.27 (2.56), CI [13.31, 27.24], F (1, 62) = 62.90, p < .001*, η_p^2 = .50.	M_{diff} = 8.09 (1.60), CI [3.72, 12.45], F (1, 62) = 25.52, p < .001*, η_p^2 = .29.
1:4	M_{diff} = 28.80 (2.91), CI [20.87, 36.72], F (1, 62) = 98.17, p < .001*, η_p^2 = .61.	M_{diff} = 9.54 (1.76), CI [4.75, 14.33], F (1, 62) = 29.43, p < .001*, η_p^2 = .32.
2:3	M_{diff} = 14.75 (2.63), CI [7.60, 21.91], F (1, 62) = 31.59, p < .001*, η_p^2 = .34.	M_{diff} = 4.43 (1.56), CI [0.17, 8.69], F (1, 62) = 8.05, p = .006, η_p^2 = .12.
2:4	M_{diff} = 23.28 (2.70), CI [15.92, 30.63], F (1, 62) = 74.35, p < .001*, η_p^2 = .61.	M_{diff} = 5.58 (1.55), CI [1.70, 10.10], F (1, 62) = 14.48, p < .001*, η_p^2 = .19.
3:4	M_{diff} = 8.52 (1.75), CI [3.75, 13.30], F (1, 62) = 23.71, p < .001*, η_p^2 = .28.	M_{diff} = 1.45 (1.33), CI [- 2.18, 5.09], F (1, 62) = 1.19, p = .28, η_p^2 = .02.

Table 3.8 Planned Comparisons for Experiment 3.
Levels column indicates the two levels that were compared. M_{diff} = difference between the means; The value in brackets is the Standard Error of the Mean; CI = 95% Confidence Intervals. Due to corrections for multiple testing, the significance criterion (p*) is .004.

Level Comparison: Paranoia * Dishonesty	
1:4 * 1:2	$F\,(1, 62) = 22.11,\ p < .001, \eta_p^2 = .26.$
1:4 * 1:3	$F\,(1, 62) = 18.86,\ p < .001, \eta_p^2 = .23.$
1:4 * 1:4	$F\,(1, 62) = 28.43,\ p < .001, \eta_p^2 = .31.$
2:4 * 1:4	$F\,(1, 62) = 11.47,\ p = .0012, \eta_p^2 = .17.$
3:4 * 1:2	$F\,(1, 62) = 20.71,\ p < .001, \eta_p^2 = .25.$
3:4 * 1:4	$F\,(1, 62) = 37.05,\ p < .001, \eta_p^2 = .37.$

Table 3.9 Post Hoc Comparisons for Experiment 3.
Significance criterion after correcting for multiple testing: $p < .0014$. All other comparisons were not significant.

Whose Rating	Target	Paranoia Rating			
		Not Paranoid	Slightly Paranoid	Moderately Paranoid	Highly Paranoid
Participant	Emily	**79.4%**	12.7%	6.3%	1.6%
	Ben	23.8%	**71.4%**	3.2%	1.6%
	Amelia	6.3%	20.6%	**66.7%**	6.3%
	George	7.9%	4.8%	22.2%	**65.1%**
Jessica	Emily	**81%**	9.5%	6.3%	3.2%
	Ben	**47.6%**	42.9%	7.9%	1.6%
	Amelia	**41.3%**	31.7%	20.6%	6.3%
	George	27%	22.2%	**30.2%**	20.6%
Oliver	Emily	**68.3%**	27%	1.6%	3.2%
	Ben	**41.3%**	39.7%	19%	0%
	Amelia	19%	**46%**	31.7%	3.2%
	George	14.3%	31.7%	**28.6%**	25.4%
Isabella	Emily	**65.1%**	17.5%	14.3%	3.2%
	Ben	39.7%	**44.4%**	15.9%	0%
	Amelia	14.3%	33.3%	**46%**	6.3%
	George	14.3%	20.6%	**34.9%**	30.2%
Jack	Emily	**60.3%**	19%	11.1%	9.5%
	Ben	**42.9%**	30.2%	19%	7.9%
	Amelia	19%	**31.7%**	28.6%	20.6%
	George	17.5%	15.9%	20.6%	**46%**

Table 3.10 Manipulation Check Frequencies for Experiment 3 (Sally Characters). Highest frequencies highlighted in bold. *N* = 63.

Whose Rating	Target	Honesty Rating			
		Not Honest	Slightly Honest	Moderately Honest	Highly Honest
Participant	Jessica	3.2%	15.9%	31.7%	**49.2%**
	Oliver	0%	22.2%	**69.8%**	7.9%
	Isabella	11.1%	**71.4%**	14.3%	3.2%
	Jack	**49.2%**	34.9%	7.9%	7.9%
Emily	Jessica	4.8%	14.3%	12.7%	**68.3%**
	Oliver	1.6%	20.6%	33.3%	**44.4%**
	Isabella	9.5%	25.4%	28.6%	**36.5%**
	Jack	15.9%	30.2%	23.8%	**30.2%**
Ben	Jessica	7.9%	19%	**42.9%**	30.2%
	Oliver	9.5%	34.9%	**52.4%**	3.2%
	Isabella	15.9%	**54%**	28.6%	1.6%
	Jack	38.1%	**39.7%**	20.6%	1.6%
Amelia	Jessica	12.7%	**44.4%**	38.1%	4.8%
	Oliver	11.1%	**57.1%**	27%	4.8%
	Isabella	38.1%	**49.2%**	11.1%	1.6%
	Jack	34.9%	**47.6%**	12.7%	4.8%
George	Jessica	**52.4%**	20.6%	17.5%	9.5%
	Oliver	**42.9%**	38.1%	17.5%	1.6%
	Isabella	**68.3%**	22.2%	6.3%	3.2%
	Jack	**68.3%**	20.6%	4.8%	6.3%

Table 3.11 Manipulation Check Frequencies for Experiment 3 (Anne characters). Highest frequencies highlighted in bold. *N* = 63.

3.11.3 Supplemental Materials for Experiment 4

Variable	*Mean*	*SD*	*Range*
Mental State Estimate:			
Sally Source Traits Level 1	83.25	22.27	8 - 100
Sally Source Traits Level 2	82.37	22.15	0 - 100
Sally Source Traits Level 3	75.82	25.07	0 - 100
Sally Source Traits Level 4	67.92	31.61	0 - 100

Table 3.12 Experiment 4: Probability of 'False Belief' Option.
Sally source traits are designed to reflect Sally's degree of paranoia if participants infer paranoia on the basis of the source traits (Level 1 = lowest, Level 4 = highest).

Whose Rating	Target	Paranoia Rating			
		Not Paranoid	Slightly Paranoid	Moderately Paranoid	Highly Paranoid
Participant	Emily	**96.36%**	3.64%	0%	0%
	Ben	32.72%	**63.64%**	3.64%	0%
	Amelia	16.36%	36.36%	**47.27%**	0%
	George	3.64%	5.46%	23.64%	**67.27%**

Table 3.13 Manipulation Check Frequencies for Experiment 4 (Sally Characters). Highest frequencies highlighted in bold. *N* = 50.

Effect	Group	Term	Estimate (95%CI)	SE	Statistic	df	*p*
Fixed	NA	(Intercept)	74.5 [69.66-, 79.36]	2.4	30.3	65	<.001**
Fixed	NA	MPRP	-8.85 [-10.65, -7.03]	0.91	-9.63	397	<.001**
Random	Participant	sd__(Interc)	15.4 [12.61, 19.01]	__	__	__	__
Random	Trial	sd__(Interc)	5.5 [4.23, 7.46]	__	__	__	__
Random	Residual	sd__Obs	20 [19.55, 20.56]	__	__	__	__

Table 3.14 Additional information for the Multiple Linear Mixed Model output.
These models assess the effect of participants' own relative placement of Sally on the paranoia dimension on the False Belief Attribution Probability. MPRP = Mean Predicted Relative Paranoia; DV = 'False Belief' Location Probability.

Model	Df	AIC	BIC	LogLik	Deviance	Chisq	Chi Df	p
Null	4	14088.92	14110.35	-7040.46	14080.92	__	__	__
Model A	5	14037.41	14064.19	-7013.70	14027.41	53.50	1	<.001**
Model B	7	14069	14106	-7027.4	14055	0.000	2	ns

Table 3.15 Model comparison for the linear mixed models in Experiment 4.
Model Null: False Belief Probability ~ (1 | trial) + (1 | participant);
Model A: False Belief Probability ~ Mean Predicted Relative Paranoia + (1 | trial) + (1 | participant);
Model B: False Belief Probability ~ Sally source traits (4 levels) + (1 | trial) + (1 | participant).

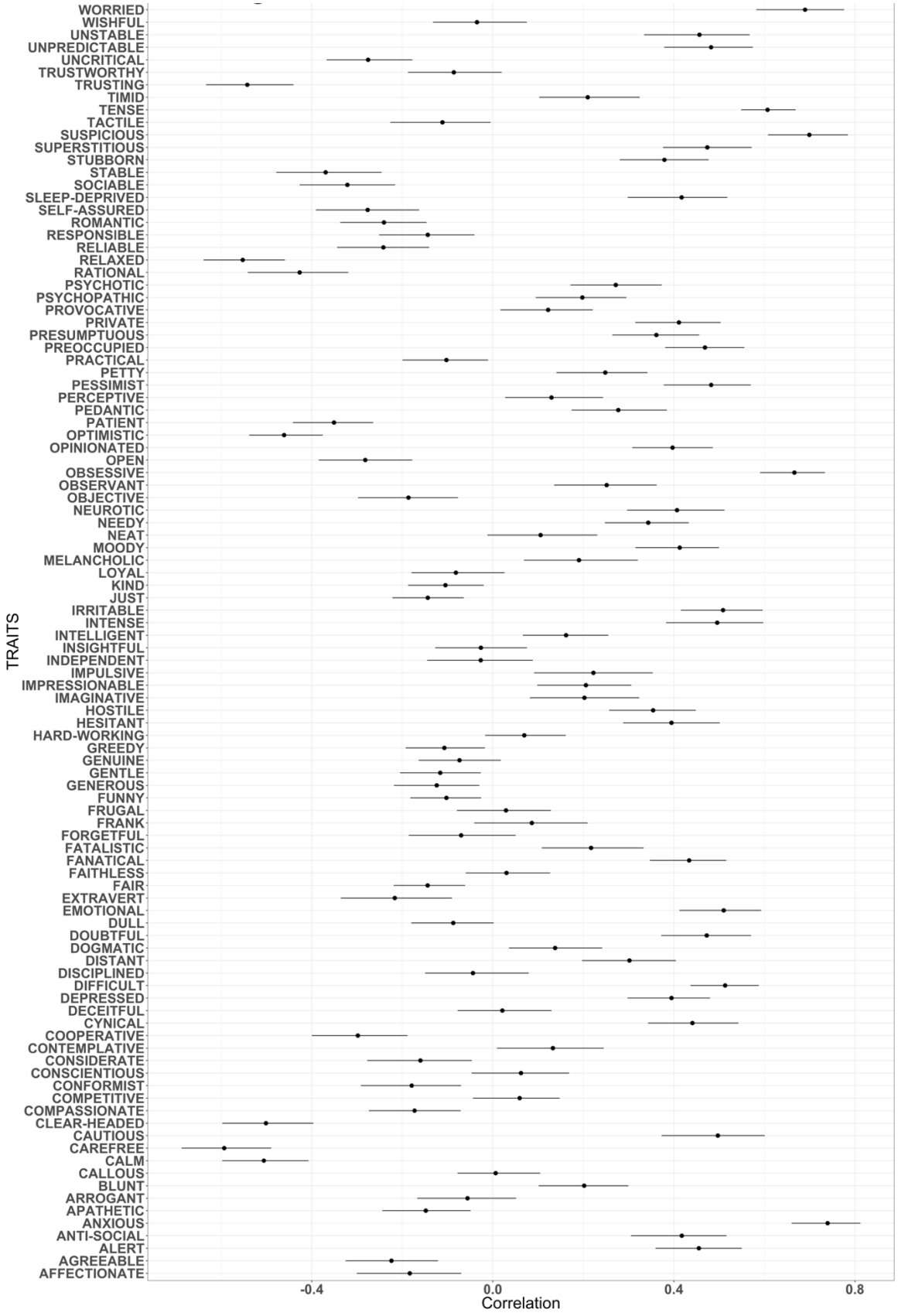

Figure 3.5 Trait correlations with paranoia obtained from the validation study for Experiment 4 (N = 50).

Data associated with Experiment 4 are provided in the file *Experiment4.csv* in the OSF server. Details about the data:

- The Sally variable is coded 1 to 4, and each level corresponds to characters Emily, Ben, Amelia and George who are designed to vary from not paranoid at all to highly paranoid.
- The MPRP (Mean Predicted Relative Paranoia) used in the LMM is computed by multiplying the weights of the Sally variable by the paranoia associations with the 6 traits (paranoia_carefree to paranoia_cautious) and then averaging those by each Sally level for each participant
- The explicit paranoia slopes used for the correlation analysis in study 4 are computed in the following manner for each participant: a) compute the mean for all the explicit paranoia ratings; b) mean centre each rating; c) add linear weights to each mean-centred rating (-3, -1, +1, +3); d) compute a sum of the weighted, mean-centred ratings. The same procedure applies for computing the MPRP slopes used in the correlation analysis for experiment 4.

NB: The computations above require switching from long to wide format in R.

3.12 References

Aboud, F. E. (1988). *Children and prejudice*. Oxford; 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. (2014). *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). New York, NY: Springer

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), 1-16.

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), 400-416.

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), 878-902.

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.

4. Adaptation of Trait Dimensions in Mind-space

4.1 Introduction

This thesis suggests that there are two main sources of variation to consider when examining mind representation and mental state inference: variation between (a) the minds available for representation, and (b) the models used by different individuals to represent those minds. The thesis links these two sources by proposing that an individual represents minds as locations in Mind-space along dimensions which allow them to be discriminated, and that the properties of individuals' Mind-spaces can differ. One component of the Mind-space framework is the mapping of minds onto dimensions, and how experience of minds in one's environment affects this process. This chapter examines whether dimensions in Mind-space adapt in response to the statistical properties of the population of experienced minds.

Adaptation is an important phenomenon that has revealed much about the neural mechanisms of perception (Rhodes et al., 2005; Webster, 2015). If we take a stimulus dimension ranging from one attribute to another (e.g. colour (Webster, 1996); tilt (Dekel & Sagi, 2015); or emotion (Skinner & Benton, 2010)), adaptation refers to a process in which prolonged exposure to one attribute biases subsequent perception towards the other attribute resulting in an 'after-effect'. For example, Skinner and Benton (2010) demonstrated that following adaptation to faces depicting six emotional anti-expressions, participants subsequently judged a neutral face as having the opposite expression to the adapting face, for instance adaptation to anti-happy resulted in a neutral face being perceived as happy. While the majority of adaptation studies have focused on the visual system (Clifford & Rhodes, 2005), evidence of after-effects has been found in other sensory modalities, including auditory (Bestelmeyer, Rouger,

DeBruine, & Belin, 2010) and haptic (van der Horst, Willebrands, & Kappers, 2008), and crossmodally (Konkle, Wang, Hayward, & Moore, 2009; Matsumiya, 2013).

Adaptation paradigms have provided empirical support for the Face-space theory that faces are represented in a multidimensional similarity space with respect to the average face at the centre (Rhodes et al., 2005; Rhodes, 2017). At a neural level, after-effects are thought to occur due to opponent coding: the average attribute activates equally a pair of neural channels, whereas above- and below- average attributes each activate one of these channels. Prolonged exposure to an adaptor stimulus therefore activates strongly one channel, which is followed by period of suppression. This suppression alters the balance of firing of the channels meaning that the average stimulus which previously caused equal activation of both channels, now results in more activation of the non-suppressed channel. Accordingly, perception is shifted away from the adaptor attribute (Rhodes, 2017).

Adaptation also demonstrates how an individual's Face-space is dynamic and malleable in response to environmental experience. Diverse environments provide very different 'sensory diets' (Webster, 2015; Webster, Werner, & Field, 2005), and adaptation effects suggest that representational systems are continuously calibrated through experience of environmental stimuli (Clifford et al., 2007; Rhodes, 2005; Webster, 2015). In lab-based experiments, the duration of exposure to the adaptor and decay of after-effects occurs over a time scale of seconds (Rhodes, Jeffery, Clifford, & Leopold, 2007). Across longer time scales, the experience-dependent calibration of Face-space is evident in the 'other-race effect', whereby individuals tend to recognize own-race faces better than other-race faces due to having more experience of own-race faces (Webster, Kaping, Mizokami, & Duhamel, 2004). Individuals who have substantial experience of other-race faces do not show such biases (Chiroro &

Valentine, 1995). Whether short-term adaptation and long-term perceptual learning are functionally similar is debated (Rhodes et al., 2005; Webster et al., 2005), but effects such as the other-race bias suggest that variation in the population of experienced faces shapes the properties of one's Face-space so that it is calibrated to best individuate between faces encountered most frequently.

The Mind-space framework proposes that, like Face-space, relative attributes of the stimulus are represented dimensionally and the space's structure is dependent on experience of the various minds in one's environment. Evidence of Mind-space after-effects would support both proposals. A challenge to the study of adaptation in Mind-space is the shift from *perceptual* to *conceptual* after-effects. Although previous adaptation studies have focused on perceptual after-effects, there is some evidence of conceptual adaptors generating perceptual after-effects. Hills et al (2010) have shown perceptual facial identity after-effects using nonvisual adaptors of voice and imagination, and when presenting a name as a written word. Furthermore, crossmodal after-effects indicate higher-level representation. For example, Matsumiya (2013) showed visual-to-haptic and haptic-to-visual after-effects for emotional facial expressions. Perceptual stimuli and abstract trait concepts have also been linked in recent work by Stolier and colleagues (Stolier, Hehman, & Freeman, 2018; Stolier, Hehman, Keller, Walker, & Freeman, 2018) in what they describe as a 'conceptual trait space'. They found that the more someone believed two personality trait concepts (e.g. caring and competent) were correlated, the more they perceived faces on those trait dimensions to be similar (Stolier et al., 2018). It is an outstanding question for 'conceptual trait space' whether conceptual adaptation would result in perceptual after-effects. In the case of Mind-space adaptation, both the adaptor and the after-effect would be conceptual rather than perceptual.

One possible empirical source of both long- and short- term conceptual adaptation is the distribution of offers in classic Game Theory studies (Lee, 2008; Güth & Kocher, 2014). For example, in the Ultimatum Game, a proposer offers a split of a given sum of money to a responder, if the responder accepts then both participants receive the sum according to the spilt offered but if the responder rejects the split then neither receive any money. The Ultimatum Game has been particularly useful for the study of fairness (Güth & Kocher, 2014; Nowak, Page, & Sigmund, 2000; Kagel, Kim, & Moser, 1996; Yamagishi et al., 2012), because a fair offer can be operationalized as 50% of the total pot. In general, proposers tend to offer 40-50% of the total amount and on average such offers are accepted (Güth & Kocher, 2014). However, there is considerable cross-cultural variation in the amount rejected by responders (Oosterbeek, Sloof, & van de Kuilen, 2004). A meta-analysis of Ultimatum Games has shown that in Bolivia and Paraguay the mean of rejected offers was 0% of the total amount, whereas in France it was 30.78% and 23.38% in the UK (Oosterbeek et al., 2004). While myriad factors may influence such cross-cultural differences, given the link between rejection rates and social norms of fairness (Kagel et al., 1996; Nowak et al., 2000), it is possible that they represent the calibration of fairness judgements to the norms exhibited in an individual's environment. Such calibration is also evident over shorter time courses. Xiang and colleagues (Xiang, Lohrenz, & Montague, 2013) assigned two groups to first receive either high or low offers in the Ultimatum Game. Both groups then received identical mid-range offers, but rejection rates and subjective ratings of emotions about the offer differed depending on experimental group, with the high-offer-adapted-group having higher rejection rates and more negative emotions.

The Xiang and cross-cultural studies suggest that concepts of fairness are responsive to the statistics of one's environment. Evidence that similar concepts are

represented dimensionally and demonstrate adaptation effects comes from fMRI adaptation studies, an experimental paradigm in which repeated presentation of a stimulus results in reduced activation relative to a novel stimulus (Ma et al., 2014; Heleven & van Overwalle, 2016). Ma et al (2014) presented participants with social traits and found adaptation effects in the ventral medial prefrontal cortex. Notably, they found adaptation not only to the same trait (e.g. honesty + honesty) but also to the opposite trait (dishonesty + honesty). The magnitude of the adaptation did not differ between the same vs. opposite trait adaptor conditions, suggesting that the high- and low- trait values are represented on the same conceptual dimension.

The current study sought to find evidence of adaptation of Mind-space dimensions in response to experience. To achieve this aim, participants were asked to act as the responder in an Ultimatum Game and rate each proposer on dimensions of modesty, generosity, and perfectionism. These three traits were chosen because modesty and generosity come from the same Honesty-Humility subscale of the HEXACO personality inventory (Ashton & Lee, 2007) as fairness, and are positively correlated. In contrast, perfectionism comes from a different subscale (Conscientiousness), is not correlated with the other traits and has no obvious relevance to the amount a proposer would offer. We predicted that experience of different distributions of adapting offers would shift the relevant trait dimensions in participants' Mind-spaces. Specifically, we predicted adaptation on dimensions of modesty and generosity but not perfectionism. A different proposer was presented on every trial to ensure that what was being adapted was the Mind-space dimension rather than the location of a particular proposer within the space. Evidence of an after-effect would be provided by traits corresponding to the same offer value being rated differently at test compared to baseline. Participants experienced two adapting conditions: High or Low offers, and offers from a Narrow or

Wide distribution. Post-adaptation, those adapted to High offers were predicted to decrease their modesty and generosity ratings whereas those adapted to Low offers were expected to increase their ratings. We predicted that a Narrow distribution would provide a stronger learning signal of the statistics of the environment, and therefore the magnitude of any after-effects would be larger in the Narrow compared to Wide conditions. Due to the correlations between the three traits, we predicted that the shift on modesty and generosity dimensions would be positively related but neither dimensional shift would relate to that on the perfectionism dimension. Finally, to test for generalisation of any after-effects we asked participants to make a charitable donation with the donation amount predicted to differ between the groups as follows: High + Narrow > High + Wide > Low + Wide > Low + Narrow.

4.2 Method

4.2.1 Participants

One hundred and twenty adults volunteered to take part in this experiment in return for a small monetary sum. Participants (33 male) were aged between 18 and 60 years old (M = 37.26, SD = 11.28). Participants were recruited via Prolific (www.prolific.ac) and the online experiment was hosted on the Gorilla platform (www.gorilla.sc; Anwyl-Irvine, Massonié, Flitton, Kirkham, & Evershed, 2019). Participants were required to be fluent in English and without a current diagnosis of a psychiatric or neurodevelopmental disorder. An *a priori* power calculation using G*Power (Faul, Erdfelder, Lang, & Buchner, 2007) indicated that for an effect size of η_p^2 = .10 and α = .05, a total sample size of 122 would provide 95% power to detect differences between the experimental groups. The local Research Ethics Committee approved the study. (Note: Two pilot versions of this experiment are detailed in the Supplemental Materials 4.5.)

4.2.2 Measures

4.2.2.1 Ultimatum Game

Participants were instructed that on every trial there was £25 to be shared between them and a Proposer, based on a split offered by the Proposer. If they accepted the proposed offer, then both they and the Proposer would receive the money according to the proposed split; if they rejected the offer then neither they nor the Proposer would receive any money. Participants were informed that there would be a new Proposer on every trial. In order for participants to play realistically, they were told in advance that both they and the Proposer would receive a bonus payment based on a randomly selected trial according to the split offered and their response on that trial.

4.2.2.2 Adaptation manipulation and trial structure

All participants completed the same Baseline and Test trials. These comprised 20 trials of offers with values randomly generated without repetition between £0 and £25. The same set of 20 trials was presented in a random order both for Baseline and Test. Between the Baseline and Test trials, participants completed 60 Adapting trials. The distributions of offers for the Adapting trials were manipulated between High and Low and Narrow and Wide. The offer values for each distribution were randomly selected from a normal distribution:

High + Narrow: M = £20 (80% of £25), SD = £1, min = £18, max = £23;
High + Wide: M = £20, SD = £2, min = £16, max = £25;
Low + Narrow: M = £5 (20% of £25), SD = £1, min = £3, max = £7;
Low + Wide: M = £5, SD = £2, min = £0, max = £9.

Participants were randomly assigned to one of the four Adapting groups.

4.2.2.3 Ratings

On all trials participants were asked to either accept or reject *the offer*. On Baseline and Test trials, participants were also asked to rate *the Proposer* on three

traits, Modesty, Generosity, and Perfectionism, on a scale from 1 (Low) to 7 (High). Participants were provided with the following definitions of each trait: Modest means not usually talking about or making obvious your own abilities and achievements; Generous means willing to give help or support, especially more than is usual or expected; Perfectionism means striving for flawlessness and setting high performance standards.

4.2.2.4 Charity Donation

All participants received a £2 bonus payment. They were asked if they would like to donate any of this bonus to charity, and to indicate the amount (from £0.00 to £2.00).

4.2.3 Procedure

Participants completed the study online in a single session of approximately 30 minutes. Participants first completed the Ultimatum Game followed by the Charity Donation.

4.3 Results

4.3.1 High vs. Low Offers

Descriptive statistics are presented in Table 4.1. For each trait, the dependent variable was the difference in rating between each of the matched 20 Test and Baseline trials, where positive scores indicate higher rating of the trait at Test, and negative scores indicate lower rating of the trait at Test. We predicted that, for Modesty and Generosity ratings, those in the High group would have negative difference scores, and those in the Low group would have positive difference scores. We predicted no change between Baseline and Test for both groups on Perfectionism ratings (a difference score of zero). To investigate these adaptation effects on trait dimensions, mixed models were

performed for each trait separately using the *lme4* package in R (Bates, Maechler, Bolker & Walker, 2018). Note that a MANOVA with the three DVs (Modesty, Generosity, Perfectionism) was not performed as the predicted correlation matrix was that only Modesty and Generosity ratings would be correlated and MANOVA is not suited to uncorrelated DVs (Tabachnick & Fidell, 2013, pp. 270). However, to control the Type I error for testing each DV separately, a Bonferroni correction was applied ($\alpha = 0.017$).

All models included the random effects, allowing for random intercepts, of Participant and Trial to account for the within-subject aspects of the design. For each trait, a baseline model (Model A) was first performed with the random effects included and without the fixed effects; then in Model B High/Low was included as a fixed effect. Model comparisons were made by comparing the change in the -2 log-likelihood values. The QQ plots of residuals showed some deviation from normality at the extremities, therefore robust regression models were also performed using the boot package in R (Canty & Ripley, 2017) to provide 95% confidence intervals of regression coefficients based on 2000 bootstrapped samples. As shown in Table 4.2, there was no significant effect of the High vs. Low manipulation on the difference scores for Modesty, Generosity, or Perfectionism ratings, and no model improved its fit compared to the baseline model.

	Modesty			Generosity			Perfectionism		
Adapting Group	M	SD	Mdn	M	SD	Mdn	M	SD	Mdn
High	0.03	0.41	0.05	-0.03	0.31	0.00	0.01	0.35	0.05
Low	0.11	0.36	0.10	0.08	0.27	0.10	-0.02	0.42	0.00
Narrow	0.03	0.43	0.00	0.05	0.29	0.05	-0.02	0.43	0.00
Wide	0.05	0.35	0.05	0.05	0.29	0.00	-0.02	0.34	-0.05

Table 4.1 Descriptive Statistics.
Means (M), standard deviations (SD), and medians (Mdn) of the 'Difference Score' dependant variable for each of the three traits and Adapting Grouping.

DV	Model	Model comparison Δ -2LL, X^2 (df), p	Predictor	B	SE (B)	95% CI	Bootstrapped 95% CI	t	p
Modesty	A		1	0.07	0.04	[-0.01, 0.14]	[-0.01, 0.14]	1.82	.07
Modesty	B	0.6, 1.32 (1), .25	High/Low	0.08	0.07	[-0.06, 0.22]	[-0.06, 0.23]	1.14	.26
Generosity	A		1	0.02	0.04	[-0.07, 0.11]	[-0.07, 0.11]	0.48	.63
Generosity	B	2, 4.00 (1), .045	High/Low	0.11	0.05	[0.00, 0.21]	[0.00, 0.21]	2.01	.047
Perfectionism	A		1	0.00	0.04	[-0.08, 0.07]	[-0.07, 0.07]	-0.08	.94
Perfectionism	B	0.1, 0.24 (1), .62	High/Low	-0.03	0.07	[-0.17, 0.10]	[-0.17, 0.10]	-0.49	.63

Table 4.2 Results for the mixed models for each trait with High vs. Low as the Fixed Effect.
***Δ -2LL* is the change in the -2 log-likelihood values. Model B was compared to Model A to test the effect on model fit of the High vs. Low predictor (compared to a constant of 1). Significance criterion = 0.017. All models included Participant and Trial as random effects.**

We predicted that High vs. Low Adapting Group would affect difference scores for Modesty and Generosity, but not Perfectionism, and the results do not support our predictions for Modesty and Generosity. As this study was exploratory, several aspects of the results are worth noting. As indicated by the large standard deviations shown in Table 4.1, there was large variation between participants in their difference scores for each trait. The means (Table 4.1) do follow the predicted order overall: difference scores are numerically greater in the Low compared to High conditions for Modesty and Generosity but not so for Perfectionism where the differences between conditions are also much smaller. Although it did not meet the significance criterion after correcting for multiple testing, with a non-corrected $\alpha = 0.05$ there was an effect of High/Low on Generosity scores (Table 4.2) and the model with the Fixed Effect did show a better fit over the baseline model. Those adapted to high offers had a lower and negative difference score ($M = -0.03$), indicating that they reduced their ratings of the Proposers' generosity, compared to those adapted to low offers who increased their generosity ratings ($M = 0.08$). The current study was sufficiently powered to detect a medium-large effect ($\eta_p^2 = .10$) at a significance criterion of 0.05; it is possible that the true effect size is much smaller and, taking into account the Bonferroni correction, a larger sample size would be required to detect any significant effects, if present.

4.3.2 Narrow vs. Wide Offers

We predicted that those in the Narrow Adapting Group would show a larger effect than those in the Wide group. In order to investigate the effect of Narrow vs. Wide Adapting Group on the trait ratings, it was necessary to reverse the signs of the difference score dependant variable described in Section 4.3.1 for those in the High group. This was due to combining the High and Low groups within the Narrow and Wide groupings; because the High group had negative predicted values and the Low

group had positive predicted values, combining them without reversing the signs of one group would result in the values cancelling each other out. Descriptive statistics are presented in Table 4.1. Analyses were performed as described previously, except now the Fixed Effect of interest was Narrow vs. Wide Adapting Group. There was no significant effect of the Narrow vs. Wide manipulation on the difference scores for Modesty, Generosity, or Perfectionism ratings, and no model improved its fit compared to the baseline model (Table 4.3). The prediction that the Narrow group would exhibit larger Modesty and Generosity after-effects than the Wide group was not supported.

DV	Model	Model comparison Δ -2LL, X^2 (df), p	Predictor	B	SE (B)	95% CI	Bootstrapped 95% CI	t	p
Modesty	A		1	0.04	0.04	[-0.03, 0.11]	[-0.03, 0.11]	1.07	.29
Modesty	B	0, 0.06 (1), .80	Narrow/Wide	0.02	0.07	[-0.12, 0.16]	[-0.12, 0.16]	0.25	.80
Generosity	A		1	0.05	0.03	[-0.01, 0.11]	[-0.01, 0.11]	1.83	0.07
Generosity	B	0, 0.002 (1), .97	Narrow/Wide	0.00	0.05	[-0.10, 0.11]	[-0.10, 0.10]	0.04	.97
Perfectionism	A		1	-0.02	0.04	[-0.09, 0.06]	[-0.09, 0.05]	-0.46	.65
Perfectionism	B	0, 0.001 (1), .97	Narrow/Wide	0.00	0.07	[-0.14, 0.14]	[-0.13, 0.13]	-0.04	.97

Table 4.3 Results for the mixed models for each trait with Narrow vs. Wide as the Fixed Effect.
***Δ -2LL* is the change in the -2 log-likelihood values. Model B was compared to Model A to test the effect on model fit of the Narrow vs. Wide predictor (compared to a constant of 1). Significance criterion = 0.017. All models included Participant and Trial as random effects.**

4.3.3 Covariance Between Trait Dimensions

To test whether difference scores on the Modesty dimension were related to those on the Generosity dimension, and neither was related to difference scores on the Perfectionism dimension, we computed a correlation matrix. As predicted, there was a significant correlation between Modesty & Generosity difference scores, $r = 0.286$, $p = .002$, but not between Modesty & Perfectionism, $r = 0.117$, $p = .20$, or Generosity & Perfectionism, $r = 0.122$, $p = .18$ (See Fig. Figure 4.1). Correlations were compared using the *concor* package in R (Lafosse, 2012). The correlation between Modesty & Generosity did not differ from that between Modesty & Perfectionism, $z = 1.45$, $p = .15$; nor did the correlation between Generosity & Perfectionism and Modesty & Perfectionism, $z = 0.05$, $p = .96$; nor between Generosity & Perfectionism and Modesty & Generosity, $z = -1.41$, $p = .16$.

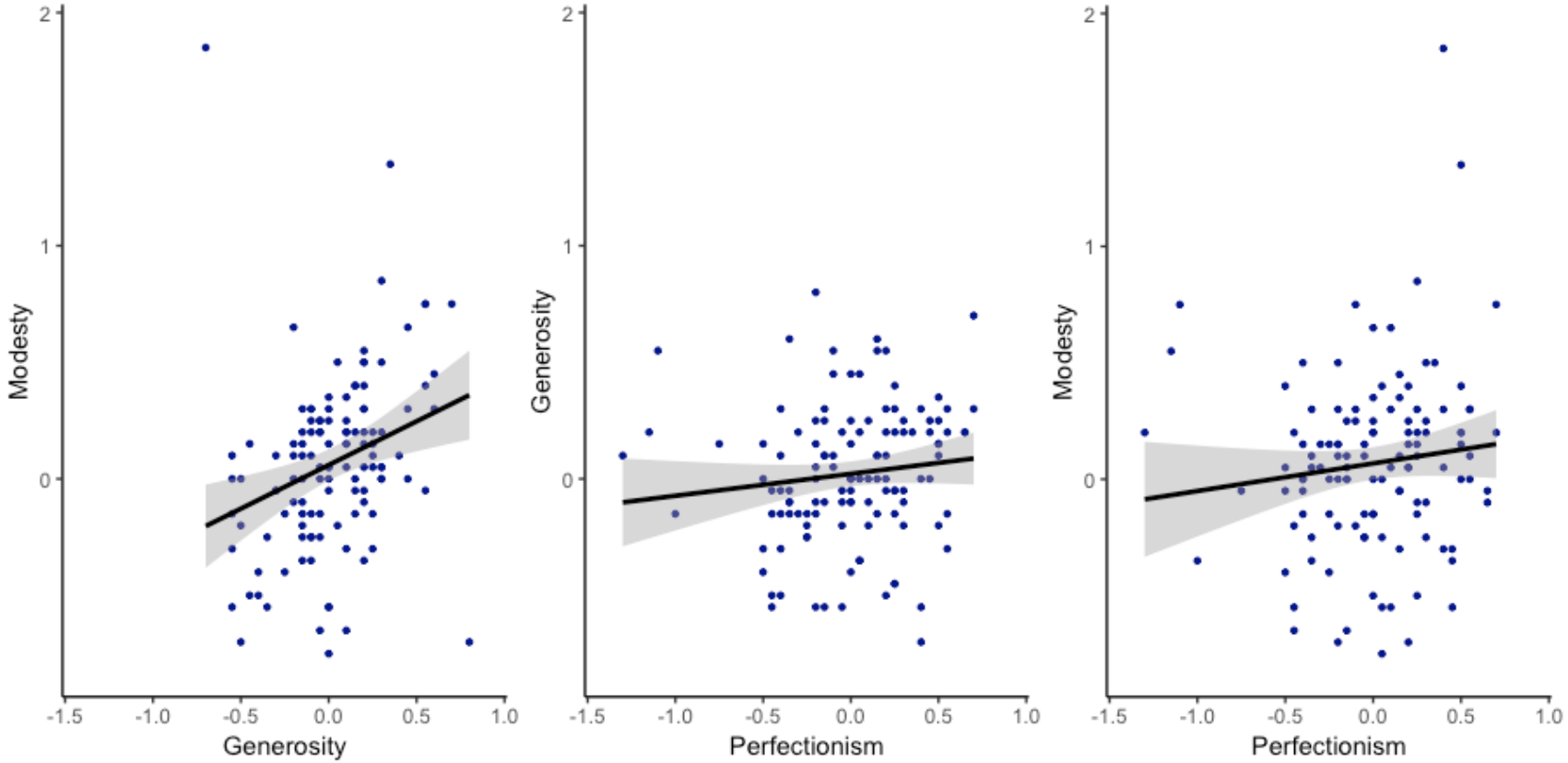

Figure 4.1 Correlations between the difference scores for each trait pair.

4.3.4 Charity Donation

To test whether Adapting Group affected participants' own levels of generosity, we compared charitable donations between groups. Due to the non-normal distribution of the data ($W = 0.72, p < .001$), a Kruskal-Wallis test was performed. There was no effect of Adaptation Group on charitable donations, $H\,(3) = 7.78, p = .05$. Although not a significant difference, as shown in Figure 4.2, the median donation was higher in the Narrow compared to Wide conditions.

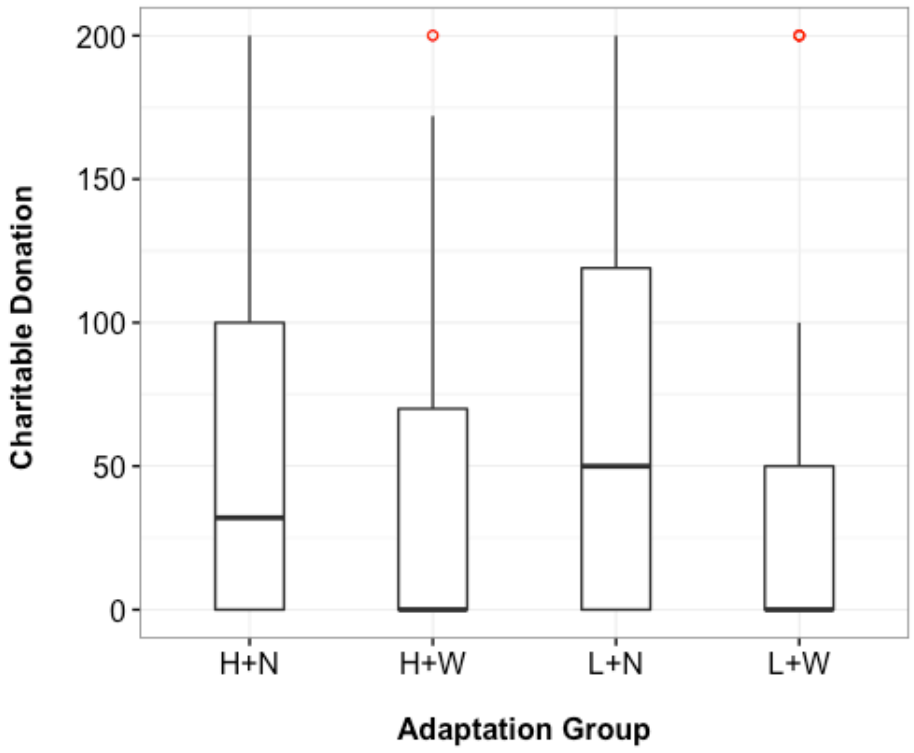

Figure 4.2 Boxplot of Charitable Donations in pence for each Adaptation Group.

4.3.5 Adaptation Effects on Acceptance of Offers

As participants' decisions whether to accept were related to the offer not the proposer, this DV was not relevant to the Mind-space hypotheses, however this test is reported for completeness. There was no effect of High vs. Low Adapting Group on acceptance difference scores between Baseline and Test, $U = 1600, p = .27$, nor was

there an effect of Narrow vs. Wide Adapting Group, $U = 1692$, $p = .57$. (Note that for comparing Narrow vs. Wide groups, the same sign reversal as described previously was used for the difference score DV.) The median of all groups was zero. This indicates that participants accepted and rejected offers similarly at Baseline and Test.

4.4 Discussion

The current study sought to find evidence of adaptation on Mind-space dimensions in response to experience of a distinct population of minds. In an Ultimatum Game, participants were adapted to high or low offers from either a narrow or wide distribution and asked to rate proposers' levels of modesty, generosity, and perfectionism. We predicted that observing after-effects for modesty and generosity, but not perfectionism, would demonstrate evidence of adaptation on Mind-space dimensions. Specifically, after-effects would indicate that the trait corresponding to the same offer value was rated differently at baseline compared to test. The direction of this difference score would indicate a shift in the trait dimension for the high vs. low offer groups, and its magnitude would show whether a narrow, compared to wide, distribution served as a stronger learning signal of population's trait levels. The findings do not support our predictions; overall, adapting group had no effect on participants' trait ratings.

As discussed in the results section, it is possible that this initial study was insufficiently powered to find effects that may be much smaller than this design could detect. For the high vs. low groups, the means did follow the predicted pattern and there is some suggestive evidence for an effect on generosity trait ratings. Taking this study and the pilots as a first exploratory step to investigate conceptual adaptation, it is possible that the experimental parameters and design were not optimal to find evidence of such. For instance, there were no trait ratings on the adapting trials and therefore it is

possible that participants did not think about the proposer on those trials but rather focused only on the offer value. Previous studies have shown that facial after-effects can occur when the adaptor involves asking the participant to visually imagine the person (Hills, Elward, & Lewis, 2010; Hills, Elward, & Lewis, 2008). Imagination adaptors produce smaller after-effects than perceptual ones (Hills et al., 2010), and they are mediated by individual differences in ability for mental visualisation (Hills et al., 2008). People also respond differently in Ultimatum Games when playing with a human compared to a computer: unfair offers were more frequently rejected from a human than a computer, and neuroimaging data showed greater activation in areas associated with the 'social brain' in response to unfair offers from a human (Sanfey, Rilling, Aronson, Nystrom, & Cohen, 2003). Given the large between-participant variation observed in the current study, future work could examine the extent to which adaptation may be mediated by the extent to which the proposer's traits are imagined. Further exploratory work could also investigate other design parameters including the number of adapting trials, the sequencing between adapting and test trials, and the strength of the adaptor.

Another result to note is the significant correlation between the shift on the generosity and modesty dimensions, and the lack of any relationship to the shift in ratings on the perfectionism dimension. As these correlations did not significantly differ from one another, this finding serves only as a starting point for future study. It suggests that experience may affect specific, related, dimensions, but not orthogonal dimensions, that the dimensions shift rather than the entire representational space. It is also of interest to determine whether participants' lay beliefs about how traits co-vary in the population underpins the strength of any correlation between dimensional shifts in Mind-space.

The results from the charitable donation did not follow the predicted pattern. The medians show that experience of a narrow compared to wide distribution resulted in more generous offers irrespective of the offer value condition being high or low. However, this result was not significant and is not readily interpretable. The finding does not support our prediction that adaptation may generalise to one's own level of generosity, which is likely attributable to the lack of any significant adaptation effect or the length of time since exposure to the adapting trials (Leopold, Rhodes, Müller, & Jeffery, 2005). A speculative explanation suggests that there may be different factors motivating giving for the high and low group. Those in the high + narrow group may have received a strong learning signal that people give generously, and knowing that others give generously has been shown to increase donations (Shang & Croson, 2009). In contrast, those in the low + narrow group may have donated for affective reasons, as giving to others has hedonic benefits (Aknin et al., 2013).

Perhaps the greatest challenge to the current study's aims was in it proposing the existence of non-perceptual after-effects. Even evidence of higher-level after-effects, such as facial identity, has been questioned as to whether these effects reflect similar mechanisms to those found using low-level non-social stimuli and whether they can indeed be called 'after-effects' (Leopold et al., 2005; Rhodes et al., 2007). The current study stated that evidence of Mind-space adaptation would be indicated by the same stimulus being rated differently post-adaptation, in effect shifting the dimension in the direction corresponding to adapting group. Clearly, this operationalisation of adaptation is not perceptual but involves the association of trait values to a visual percept of a monetary sum. Previous studies using the Ultimatum Game have emphasised the learning of social-norms (Xiang et al., 2013) and the violation of social expectations (Chang & Sanfey, 2013), but have not shown how experience may affect how the proposer, rather than the offer, is represented. A further challenge to the idea of

conceptual adaptation is whether any effects would in fact reflect a change in a decision-making heuristic, or criterion (Storrs, 2015), rather than a change in how someone is actually represented in Mind-space.

In conclusion, the current study drew on the analogy between Face-space and Mind-space and explored whether there was evidence of adaptation on trait dimensions in Mind-space. The findings indicate that experience of different distributions of offers in the Ultimatum Game had no effect on trait ratings of the proposer. Future investigations of whether and how Mind-space dimensions recalibrate over the short- and long- term may require significantly different designs to those utilised in the perceptual literature.

4.5 Supplemental Materials

4.5.1 Pilot Experiments

Two pilot experiments were run prior to the current study. The initial experimental design was based more closely on psychophysics studies demonstrating adaptation after-effects in sensory modalities. In such studies, stimuli are created along a dimension, as seen in Figure 4.3, ranging from one attribute of the stimulus (A) to another (B) in a series of steps representing the blending of the two attributes. For instance, attributes may reflect emotional facial expressions of anger to fear (Bestelmeyer et al., 2010), or the directional tilt of a shape from left to right (Dekel & Sagi, 2015). Participants are asked to select the attribute of a presented stimulus from the two alternatives (A vs. B; a 'two-alternative forced-choice'). This method allows a psychometric function to be fitted for each participant that describes the relationship between the probabilities of the participant's response (A vs. B) and the stimulus blending steps (Kingdom & Prins, 2010). The 'Point of Subjective Equality' (Kingdom & Prins, 2010, pp. 265) can then be derived; this reflects the relative percentage of each

attribute at a blending step where the participant was equally likely to rate the stimulus as having either attribute.

Adaptation through exposure to one extreme attribute of the stimulus (100% A) results in biased perception towards the alternative attribute (B). For example, following adaptation to 100% male faces, faces rated prior to adaptation as of neutral gender were now rated as female, and the point at which faces were rated as neutral occurred closer to the male end of the dimension (Webster et al., 2004). These after-effects demonstrate that adaptation can shift the PSE, and dimensions in participants' face-spaces, based on stimuli recently experienced. These pilot studies aimed to use a similar two-alternative forced-choice adaptation paradigm to test whether exposure to different distributions of offers in an Ultimatum Game shifted the PSE on trait dimensions, thus demonstrating that experience affects individuals' representations of Mind-space and the dimensional therein.

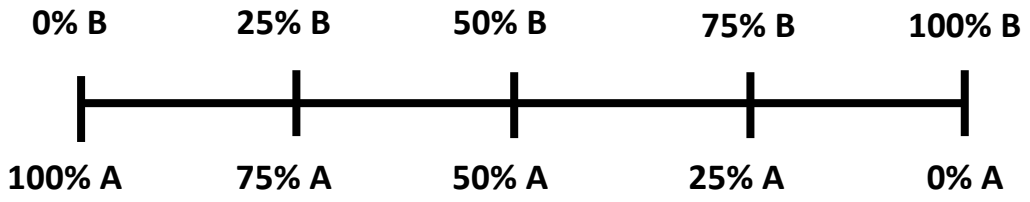

Figure 4.3 An example of stimulus blending steps (expressed as a percentage) from one attribute (A) to another (B).

4.5.1.1 Methods

As in the current study, participants played the Ultimatum Game, there were four Adapting Groups (High + Narrow; High + Wide; Low + Narrow; Low + Wide), and the Mind-space dimensions were Modesty, Generosity, and Perfectionism. A

continuum of offers from low to high was created with five test stimulus steps (conceptually similar to the stimulus blending steps shown in Figure 4.3). In Pilot 1, the total range of possible offers from Proposers was between £0 and £24, and the test stimulus steps were: £4.70, £8.35, £12, £15.65, and £19.30. In Pilot 2, the total range of possible offers was 0 to 240 points (which corresponded to monetary value), and the test stimulus steps also included a jitter around each step: 45/50/55 points, 80/85/90 points, 115/120/125 points, 150/155/160 points, and 185/190/195 points. The distributions of the four Adapting Groups are shown in Table 4.4. In Pilot 1, the adapting offers overlapped with the range of the test stimulus steps, whereas in Pilot 2 the adapting offers were outside the range.

	Pilot 1		Pilot 2	
Adapting Group	M	SD	M	SD
High + Narrow	£19	£0.25	225 points	1 points
High + Wide	£19	£1.50	225 points	5 points
Low + Narrow	£5	£0.25	15 points	1 points
Low + Wide	£5	£1.50	15 points	5 points

Table 4.4 Distributions of the offers for each Adapting Group in pilot studies M = Mean, SD = Standard Deviation.

Participants first completed 30 Baseline trials comprising each of the five test stimulus levels presented six times in a random order. This was followed by six blocks of trials. The first ten trials of each block were adapting trials, followed by a test trial, then a pattern of four adapting trials followed by a test trial which was repeated a further three times. Therefore each block had 26 adapting trials and five test trials (presented in a random order). On each trial participants were presented with the Proposer's offer. On all trials they made a two-alternative forced-choice response to *the offer* of Accept or Reject. On all test trials, they also made three further two-alternative

forced-choice responses, relating to *the Proposer*, of: Modest or Immodest; Generous or Greedy; and Perfectionist or Not a Perfectionist.

The intended dependent variable was the Point of Subjective Equivalence (PSE). We hypothesised that the PSE would shift between Baseline and Test according to Adapting Group for Modesty and Generosity, but not Perfectionism. Specifically, that the PSE would move to a higher value for the High Adapting Group, and to a lower value for the Low Adapting Group, and that this pattern would be more pronounced in the Narrow compared to Wide Adapting subgroups. For example, if a participant's PSE for Generous vs. Greedy occurred at £12 (50% split with the Proposer) at Baseline, and they were then adapted to a High and Narrow distribution of offers, we predicted that their PSE would shift to a higher value, i.e. their PSE for Generous vs. Greedy might now occur at £15.65, meaning that their threshold for rating someone as generous had increased. Psychometric functions were fitted using the Palamedes Matlab toolbox (Prins & Kingdom, 2018). However, in both pilots (P1: $N = 8$; P2: $N = 4$) it was not possible to fit a psychometric curve from which to derive a PSE for any participant, and therefore not possible to analyse the data as planned. Examples of PSE fits are shown in Figure 4.4.

In these pilot studies the number of trials at each test stimulus levels was small compared to trial numbers in the perceptual adaptation literature (Kingdom & Prins, 2010, pp. 62). This was in order to keep the total experiment time to a minimum (max 30 minutes) to keep participants engaged in the task given its repetitive nature. There were only 6 trials at each test stimulus level. In combination with the two-alternative forced-choice response type, this resulted in the probability dimension (the y-axis of the graphs in Figure 4.4) having a poor resolution. In general, participants either made the same response at all stimulus steps (e.g. always or never rating the Proposer as

generous) or exhibited a binary strategy whereby all offers over a certain value were always rated as generous. It is also possible that the number of adapting trials was insufficient. For example, the Xiang et al study (2013) used 30 consecutive trials of offers for social norm training in the Ultimatum Game. Therefore, the experiment was redesigned to the version presented in the current study.

Figure 4.4 a Reject vs. Accept

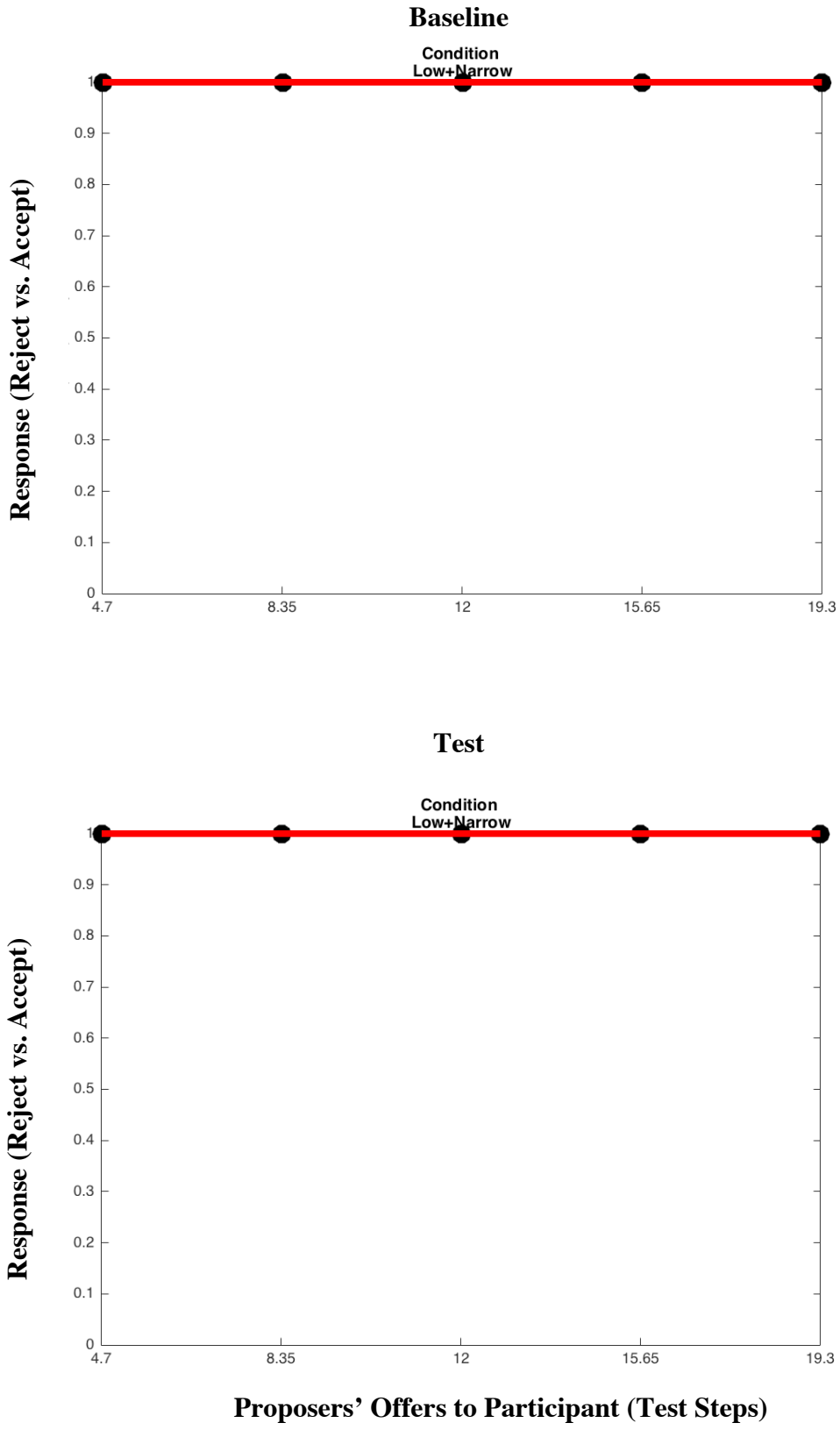

Figure 4.4 b Greedy vs. Generous

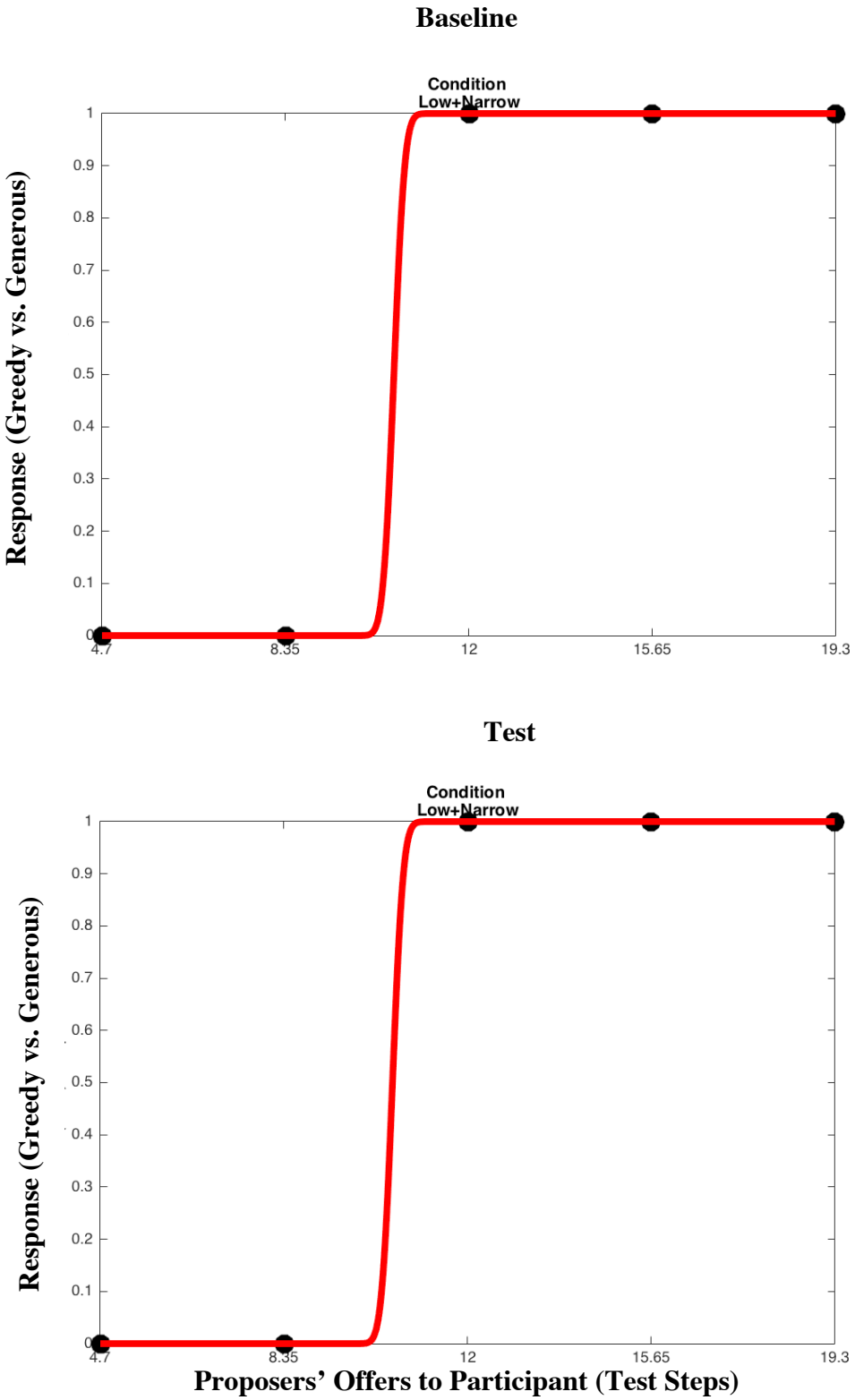

Figure 4.4 c Immodest vs. Modest

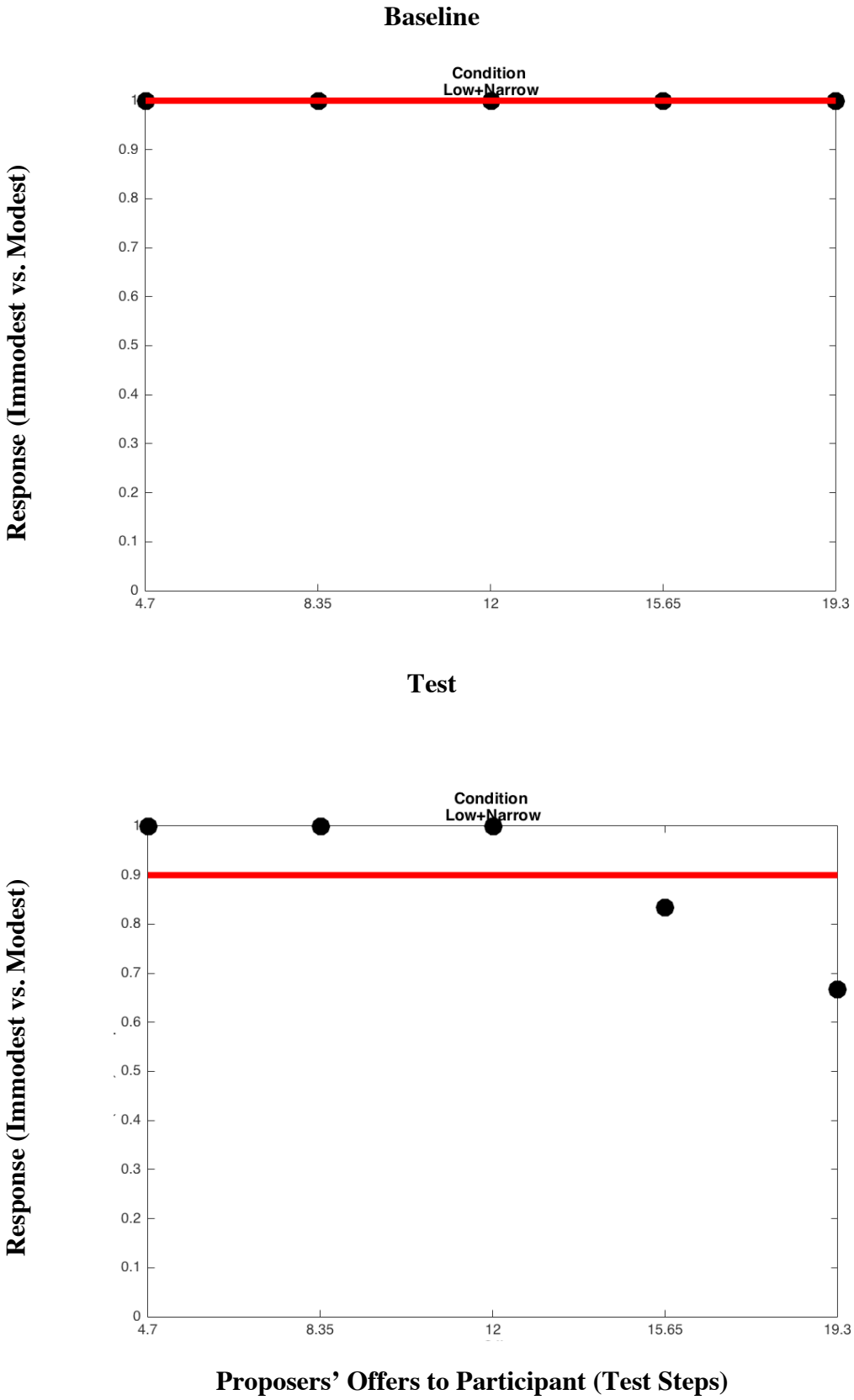

Figure 4.4 d Not a Perfectionist vs. Perfectionist

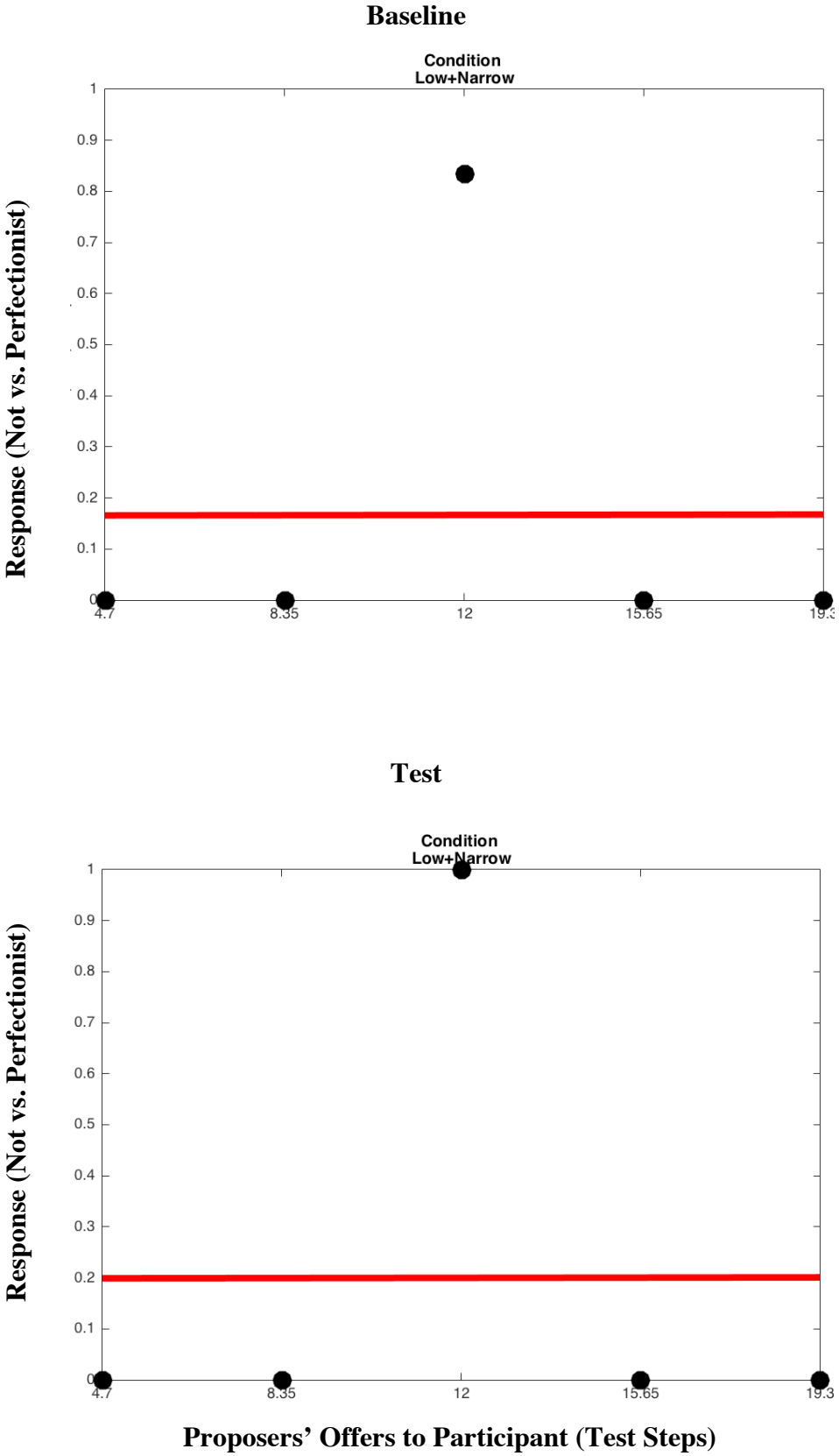

Figure 4.4 Psychophysical functions for the Baseline and Test Conditions for one pilot participant.
Responses show: (a) Acceptance; (b) Generosity; (c) Modesty; (d) Perfectionism. This participant was in the Low + Narrow offers condition.

4.6 References

Aknin, L. B., Barrington-Leigh, C. P., Dunn, E. W., Helliwell, J. F., Burns, J., Biswas-Diener, R., … Norton, M. I. (2013). Prosocial spending and well-being: Cross-cultural evidence for a psychological universal. *Journal of Personality and Social Psychology*, *104*(4), 635–652. https://doi.org/10.1037/a0031578

Anwyl-Irvine, A.L., Massonié J., Flitton, A., Kirkham, N.Z., Evershed, J.K. (2019). Gorilla in our midst: an online behavioural experiment builder. *Behavior Research Methods*. doi: https://doi.org/10.3758/s13428-019-01237-x

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

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/

Bestelmeyer, P. E. G., Rouger, J., DeBruine, L. M., & Belin, P. (2010). Auditory adaptation in vocal affect perception. *Cognition*, *117*(2), 217–223. https://doi.org/10.1016/j.cognition.2010.08.008

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

Chang, L. J., & Sanfey, A. G. (2013). Great expectations : neural computations underlying the use of social norms in decision-making. *Social Cognitive and Affective Neuroscience*, *8*, 277–284. https://doi.org/10.1093/scan/nsr094

Chiroro, P., & Valentine, T. (1995). An investigation of the contact hypothesis of the

own-race bias in face recognition. *The Quarterly Journal of Experimental Psychology*, *48*(4), 879–894. https://doi.org/10.1080/14640749508401421

Clifford, C. W. G., & Rhodes, G. (2005). *Fitting the Mind to the World: Adaptation and After-Effects in High-Level Vision*. Oxford, UK: Oxford University Press

Clifford, C. W. G., Webster, M. A., Stanley, G. B., Stocker, A. A., Kohn, A., Sharpee, T. O., & Schwartz, O. (2007). Visual adaptation : Neural, psychological and computational aspects. *Vision Research, 47*, 3125–3131. https://doi.org/10.1016/j.visres.2007.08.023

Dekel, R., & Sagi, D. (2015). Tilt aftereffect due to adaptation to natural stimuli. *Vision Research*, *117*, 91–99. https://doi.org/10.1016/j.visres.2015.10.014

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

Güth, W., & Kocher, M. G. (2014). More than thirty years of ultimatum bargaining experiments: Motives, variations, and a survey of the recent literature. *Journal of Economic Behavior and Organization*, *108*, 396–409. https://doi.org/10.1016/j.jebo.2014.06.006

Heleven, E., & Van Overwalle, F. (2016). The person within: Memory codes for persons and traits using fMRI repetition suppression. *Social Cognitive and Affective Neuroscience*, *11*(1), 159–171. https://doi.org/10.1093/scan/nsv100

Hills, P. J., Elward, R. L., & Lewis, M. B. (2008). Identity adaptation is mediated and moderated by visualisation ability. *Perception*, *37*(8), 1241–1257. https://doi.org/10.1068/p5834

Hills, P. J., Elward, R. L., & Lewis, M. B. (2010). Cross-modal face identity aftereffects and their relation to priming. *Journal of Experimental Psychology: Human Perception and Performance*, *36*(4), 876–891. https://doi.org/10.1037/a0018731

Kagel, J. H., Kim, C., & Moser, D. (1996). Fairness in ultimatum games with asymmetric information and asymmetric payoffs. *Games and Economic Behavior*, *13*(1), 100–110. https://doi.org/10.1006/game.1996.0026

Kingdom, F.A.A. & Prins, N. (2010). *Psychophysics: A Practical Introduction*. Oxford, UK: Academic Press, Elsevier.

Lafosse, R. (2012). *Concor*. Retrieved from https://cran.r-project.org/web/packages/concor/concor.pdf.

Lee, D. (2008). Game theory and neural basis of social decision making. *Nature Neuroscience*, *11*(4), 404–409.

Leopold, D. A., Rhodes, G., Müller, K. M., & Jeffery, L. (2005). The dynamics of visual adaptation to faces. *Proceedings of the Royal Society B: Biological Sciences*, *272*(1566), 897–904. https://doi.org/10.1098/rspb.2004.3022

Ma, N., Baetens, K., Vandekerckhove, M., Kestemont, J., Fias, W., & Van Overwalle, F. (2014). Traits are represented in the medial prefrontal cortex: An fMRI adaptation study. *Social Cognitive and Affective Neuroscience*, *9*(8), 1185–1192. https://doi.org/10.1093/scan/nst098

Matsumiya, K. (2013). Seeing a Haptically Explored Face: Visual Facial-Expression Aftereffect From Haptic Adaptation to a Face. *Psychological Science*, *24*(10), 2088–2098. https://doi.org/10.1177/0956797613486981

Nowak, M. A., Page, K. M., & Sigmund, K. (2000). Fairness Versus Reason in the

Ultimatum Game. *Science*, *289*(5485), 1773.

Oosterbeek, H., Sloof, R., & van de Kuilen, G. (2004). Cultural Differences In Ultimatum Game Experiments: Evidence From A Meta-Analysis. *Experimental Economics*, *7*, 171–188. https://doi.org/10.2139/ssrn.286428

Prins, N & Kingdom, F. A. A. (2018) Applying the Model-Comparison Approach to Test Specific Research Hypotheses in Psychophysical Research Using the Palamedes Toolbox. Frontiers in Psychology, 9:1250. doi: 10.3389/fpsyg.2018.01250

Prolific, (2014). Available at: https://www.prolific.ac

Rhodes, G. (2017). Adaptive Coding and Face Recognition. *Current Directions in Psychological Science*, *26*(3), 218–224. https://doi.org/10.1177/0963721417692786

Rhodes, G., Jeffery, L., Clifford, C. W. G., & Leopold, D. A. (2007). The timecourse of higher-level face aftereffects. *Vision Research*, *47*(17), 2291–2296. https://doi.org/10.1016/j.visres.2007.05.012

Rhodes, G., Robbins, R., Jaquet, E., Mckone, E., Jeffery, L., & Clifford, C. W. G. (2005). Adaptation and Face Perception: How Aftereffects Implicate Norm-Based Coding of Faces. In C. W. G. Clifford & G. Rhodes (Eds.), *Fitting the Mind to the World Adaptation and After-Effects in High-Level Vision* (pp. 213–241). Oxford: Oxford University Press. https://doi.org/10.1093/acprof:oso/9780198529699.003.0009

Sanfey, A. G., Rilling, James, K., Aronson, J. A., Nystrom, L. E., & Cohen, Jonathon, D. (2003). The Neural Basis of Economic Decision-Making in the Ultimatum

Game. *Science*, *300*, 1755–1759. https://doi.org/10.1126/science.1082976

Shang, J., & Croson, R. (2009). A Field Experiment In Charitable Contribution : The Impact Of Social Information On The Voluntary Provision Of Public Goods. *The Economic Journal*, *119*, 1422–1439. https://doi.org/10.1111/j.1468-0297.2009.02267.x.

Skinner, A. L., & Benton, C. P. (2010). Anti-expression aftereffects reveal prototype-referenced coding of facial expressions. *Psychological Science*, *21*(9), 1248–1253. https://doi.org/10.1177/0956797610380702

Stolier, R. M., Hehman, E., & Freeman, J. B. (2018). A Dynamic Structure of Social Trait Space. *Trends in Cognitive Sciences*, *22*(3), 197–200. https://doi.org/10.1016/j.tics.2017.12.003

Stolier, R. M., Hehman, E., Keller, M. D., Walker, M., & Freeman, J. B. (2018). The conceptual structure of face impressions. *Proceedings of the National Academy of Sciences*, *115*(37), 9210–9215. https://doi.org/10.1073/pnas.1807222115

Storrs, K. R. (2015). Are high-level aftereffects perceptual? *Frontiers in Psychology*, *6*(February), 6–9. https://doi.org/10.3389/fpsyg.2015.00157

Tabachnick, B. G., & Fidell, L. S. (2013). *Using Multivariate Statistics* (5th ed., pp. 270). Boston, MA: Pearson

van der Horst, B. J., Willebrands, W. P., & Kappers, A. M. L. (2008). Transfer of the curvature aftereffect in dynamic touch. *Neuropsychologia*, *46*(12), 2966–2972. https://doi.org/10.1016/j.neuropsychologia.2008.06.003

Webster, M. A. (1996). Human colour perception and its adaptation. *Network: Computation in Neural Systems*, *7*(4), 587–634. https://doi.org/10.1088/0954-

898X_7_4_002

Webster, M. A. (2015). Visual Adaptation. *Annual Review of Vision Science*, *1*, 547–567. https://doi.org/10.1146/annurev-vision-082114-035509

Webster, M. A., Kaping, D., Mizokami, Y., & Duhamel, P. (2004). Adaptation to natural facial categories. *Nature*, *428*(April), 557–561. https://doi.org/10.1038/nature02361.1.

Webster, M. A., Werner, J. S., & Field, D. J. (2005). Adaptation and the Phenomenology of Perception. In C. W. G. Clifford & G. Rhodes (Eds.), *Fitting the Mind to the World: Adaptation and After-Effects in High-Level Vision*, (pp. 241-278). Oxford: Oxford University Press. https://doi.org/10.1093/acprof:oso/9780198529699.003.0009

Xiang, T., Lohrenz, T., & Montague, P. R. (2013). Computational substrates of norms and their violations during social exchange. *The Journal of Neuroscience*, *33*(3), 1099–1108. https://doi.org/10.1523/JNEUROSCI.1642-12.2013

Yamagishi, T., Horita, Y., Mifune, N., Hashimoto, H., Li, Y., Shinada, M., … Simunovic, D. (2012). Rejection of unfair offers in the ultimatum game is no evidence of strong reciprocity. *Proceedings of the National Academy of Sciences*, *109*(50), 20364–20368. https://doi.org/10.1073/pnas.1212126109

5. The Role of the Self and Metacognition in Modelling Another's Memory

5.1 Introduction

Despite important commonalities between research on the representation of minds and of mental states, metacognition and theory of mind have tended to be studied separately, forming distinct literatures with distinct testing methods (Flavell, 2000). Metacognition refers to 'cognition about cognition', meaning a representation whose object is the property of a cognitive process (Flavell, 2000; Shea et al., 2014), and, in practice, studies have focused on the monitoring and control of one's own cognitive processes, such as memory (Dunlosky & Tauber, 2015), within a task or with respect to a goal (Nelson, Narens, & Dunlosky, 2004). In contrast, theory of mind tends to refer to one's representation of another person's mental representation ('metarepresentation'; Leslie, 1987), and, as outlined in Chapter 2 (2.4), studies have mostly focused on whether one can correctly infer another's mental state.

One link between these two literatures is empirical evidence showing that those with autism spectrum disorder show impairments when making inferences about their own mental states as well as those of others (Frith & Happé, 1999; Williams, 2010), and perform poorly on both metacognitive and theory of mind tasks (Grainger, Williams, & Lind, 2014). These observations support the idea that a theory of other minds relies on the same mechanisms as a theory of own mind - in effect, metacognition. Indeed, representations of one's own mental states have long been considered to bias (Epley, Keysar, van Boven, & Gilovich, 2004), or even underpin (Perner &Kuhlberger, 2005), representations of another's mental states.

To better bridge research on metacognition and theory of mind, it is necessary to

outline the relationship between metarepresentations of cognitive processes and of mental states. The Mind-space framework makes a distinction between minds and mental states, specifically that minds are the complete set of cognitive systems, and mental states are the representational content generated by that set of systems; minds are represented along dimensions in Mind-space and the location of a mind relates to a probability of it having a particular mental state. Therefore, a dimension may represent memory and the location of a target mind on that dimension would affect the mental state attributed to it. As described in Figure 2.3 in Chapter 2, a forgetful person would likely have different beliefs about the location of an object compared to someone with a good memory.

Within the Mind-space framework, metacognitive accuracy can be described as the ability to locate the self in Mind-space. If one locates accurately the self and similar others in the space, then one's own mind serves as a good model for inferring their mental states, due to the use of the same mapping of location in Mind-space to mental state probability. Conversely, if one has poor metacognitive accuracy then the self is no longer a good model for other minds and their mental states. The Mind-space framework therefore suggests a more detailed explanation for *why* metacognition of one's own mind may affect representations of the self's and others' mental states, and *when* the self is a good model for others. Of further consideration is whether the self is used as a model for a mind about whom one knows nothing, and whether the location of the self on certain dimensions affects the space architecture, for example one's trait self-esteem may bias the locating of others in relation to the self. (These points are discussed in Section 2.8.)

The current study sought to examine the representation of another's memory, how this is affected by one's own memory and by how accurately one represents one's

own memory. To achieve this aim, subjects were introduced to two memory tasks by performing them as single tasks, and then asked to estimate another participant's performance when completing the two tasks as a dual task. (Note that for clarity, the people who actually took part in the current study are referred to as subjects, and the person the subject was asked about is referred to as a participant). Subjects were given no details about the other participant. Subjects were then given specific experience of the dual task themselves, and asked once again to estimate another participant's dual task performance. This design allowed us to measure how subjects modelled another's decrement in performance across the dual task; whether subjects used their own performance to predict another's, and if this depended on their metacognitive accuracy; and how experience affected predictions. Finally, the subject's theory of mind ability, trait self-esteem, and perspective-taking tendencies were also examined to investigate whether they affected how they modelled another's memory.

5.2 Method

5.2.1 Subjects

Sixty-seven adults volunteered to take part in this experiment in return for a small monetary sum or undergraduate research participation credits. Subjects (9 male) were aged between 18 and 32 years old (M = 19.76, SD = 2.66). 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 66 would provide 80% power for the main hypothesis being tested (with three predictor variables). The local Research Ethics Committee approved the study.

5.2.2 Measures

5.2.2.1 Task A: Letter Span Task

In this task a series of letters was presented on screen, one letter at a time. There were four levels of increasing task difficulty. The numbers of letters in each series varied according to the difficulty level of Task A: at Level 1 there were three letters in the series, four letters at Level 2, five letters at Level 3, and six letters at Level 4. The letter series were designed so they did not spell a word. Each letter series constituted one trial, and seven trials were presented at each level. At the end of every trial, subjects were asked to recall the letter series in the correct order within a max response time of six seconds. A correct response required that the full series be recalled in the correct order. The subject's percentage accuracy at each task level was computed. In addition, at the end of each task level, subjects were asked to estimate the percentage of trials on which they responded correctly. The subject's actual accuracy was subtracted from their estimated accuracy to provide a measure of metacognitive accuracy.

5.2.2.2 Task B: N-back Task

In this task a series of six numbers was presented on screen, one number at a time. Within this series, sometimes the same number appeared as N numbers previously (e.g. a 2-back: 4 - 1 - 4). There were four levels of increasing difficulty. The N of the target N-back varied according to the difficulty level of Task B: a 1-back at Level 1, a 2-back at Level 2, a 3-back at Level 3, and a 4-back at Level 4. Each number series constituted one trial, and seven trials were presented at each level; 50% of all trials featured the target N-back for that level, 20% had no N-back, and 30% had a non-target N-back. At the end of every trial, subjects were asked to report whether the target N-back had occurred (yes/no) within a max response time of six seconds. The subject's percentage accuracy at each task level was computed. As for Task A, at the end of each

task level, subjects were asked to estimate the percentage of trials on which they responded correctly. The subject's actual accuracy was subtracted from their estimated accuracy to provide a measure of metacognitive accuracy.

5.2.2.3 Tasks A and B as a Dual Task

As a dual task, both the letter series of Task A and the number series of Task B were presented simultaneously onscreen in a vertical array. Which task appeared above the other was randomised across subjects. Subjects completed every level of Task A at every level of Task B, and the order of these 16 level combinations was presented randomly. Subjects responded as previously described for each task, this means that at each of the 16 levels of the dual task, subjects responded to Task A and to Task B. The measure for dual task accuracy was computed as the average accuracy at each level combination ((Task A + Task B % Accuracy)/2). Three metacognitive accuracies were computed: average Task A metacognitive accuracy; average Task B metacognitive accuracy; and average overall metacognitive accuracy. These metacognitive accuracy measures were mean-centred to serve as moderating variables in the analyses.

5.2.2.4 Judging Another Participant's Performance on the Dual Task

Subjects were told that another participant completed Tasks A and B as a dual task, and their accuracy and metacognitive accuracy was measured for each of the 16 level combinations. Subjects were asked to estimate:

1. On what percentage of trials did the other participant respond correctly;
2. On what percentage of trials did the other participant think that they respond correctly.

These measures were labelled as 1st and 2nd order Participant Accuracy Estimate respectively, and were recorded for each of the 16 level combinations for Task A and for Task B.

5.2.2.5 Theory of Mind Measure

The Movie for the Assessment of Social Cognition (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 subject's scores were computed as the percentage of correct responses on the mental state questions.

5.2.2.6 Trait Measures.

Subjects also completed 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, and the Rosenberg Self-esteem Scale (Gray-Little, Williams, & Hancock, 1997).

5.2.3 Procedure

Subjects first completed Task A and Task B as single tasks, to familiarize themselves with the tasks. Following this they were asked to judge another participant's performance on Tasks A and B when performed as a dual task. Subjects then completed Tasks A and B as a dual task themselves, and were then asked for a second time to judge another participant's performance on Tasks A and B as a dual task. Finally, subjects completed the MASC and trait measures. Subjects completed the study individually on a computer in a testing room in a single session of approximately two hours.

5.2.4 Data-preprocessing

The key experimental manipulation in this study is the increasing difficulty of Tasks A and B, which is predicted to affect performance on the dual task. The dual task measures, described earlier, provide the following dependent variables:

- Subject's Accuracy;
- Subject's Metacognitive Accuracy;
- 1st Order Participant Accuracy Estimate - Before Dual Task Experience;
- 2nd Order Participant Accuracy Estimate - Before Dual Task Experience;
- 1st Order Participant Accuracy Estimate - After Dual Task Experience;
- 2nd Order Participant Accuracy Estimate - After Dual Task Experience.

Subject's Metacognitive Accuracy was computed as described earlier and served as a moderating variable. For the other five variables, in order to quantify how the increasing difficulty of Tasks A and B in the dual task affected them, 3D regression models were performed for each variable separately to generate slope coefficients for each subject. These slope coefficients quantify the degree to which each subject's dual task variables were affected by the increasing levels of Task A, Task B, and their interaction Tasks A x B (i.e. both tasks rising together). The regression model for each of the five dual task dependent variables (y) listed above was as follows: y ~ Task A Level + Task B Level + Tasks A x B Levels.

For example, for the dependent variable of Subject's Accuracy, a subject's Task A regression coefficient represents the degree to which that the subject's accuracy on the dual task was affected by increasing levels of Task A (note that the accuracy measure refers to dual task accuracy, not Task A only accuracy); similarly, the subject's Task B regression coefficient represents the degree to which that subject's accuracy on

the dual task was affected by increasing levels of Task B. The A x B coefficient represents the degree to which that subject's accuracy on the dual task was affected by increasing levels of Tasks A and B together, beyond the effect that was accounted for by the Task A and Task B coefficients.

Figure 5.1 illustrates the interpretation of a 3D regression and a Tasks A x B interaction, which is represented by a surface. Figure 5.1a shows a slope of -9 for Task A and -12 for Task B, but a zero A x B interaction. This means that the accuracy at A4B4 is predicted only by the degree to which accuracy is affected by Task A (-9% per level) and Task B (-12% per level), therefore accuracy at A4B4 = 37% [(-9 x 3) + (-12 x 3) = -63; A1B1 = 100, A4B4 = 100 - 63 = 37]. A zero A x B interaction is represented by a flat surface. In contrast, Figure 5.1b shows the same slopes for Task A and Task B, but with a negative A x B interaction. This means that the accuracy at A4B4 is predicted not only by the degree to which accuracy is affected by Task A and Task B but also by the A x B interaction, which in this example is -4, therefore accuracy at A4B4 = 1% [(-9 x3)+(-12 x 3)+(-4 x 3 x 3) = -99; A1B1 = 100, A4B4 = 100-99 = 1%]. A negative A x B interaction is represented by the surface bending further downwards at the A4B4 corner (compared to Figure 5.1a), and indicates that accuracy was additionally affected by the levels of Tasks A and B rising together. Figure 5.1c shows the same slopes for Task A and Task B, but with a positive A x B interaction. In this example the A x B interaction is + 4, therefore accuracy at A4B4 = 73% [(-9 x3)+(-12 x 3)+(+4 x 3 x3) = -27; A1B1 = 100, A4B4 = 100-27 = 73]. A positive A x B interaction is represented by the surface rising upwards at the A4B4 corner, and indicates that accuracy was not as affected by the levels of Task A and B rising together as would be predicted by the slopes of Task A and Task B.

Figure 5.1 a **A x B Interaction = 0**

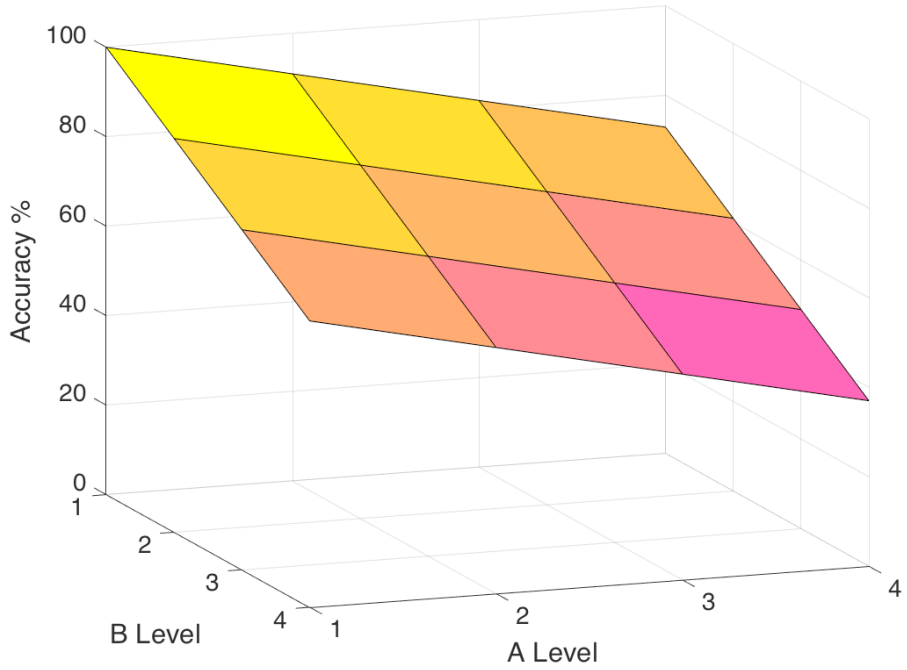

Figure 5.1 b **A x B Interaction is Negative**

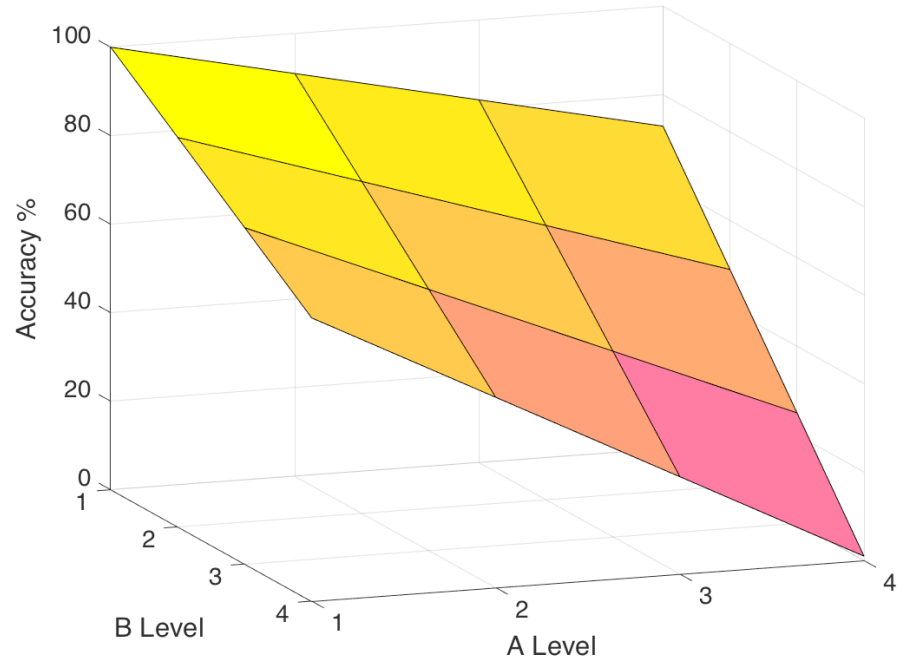

Figure 5.1 c A x B Interaction is Positive

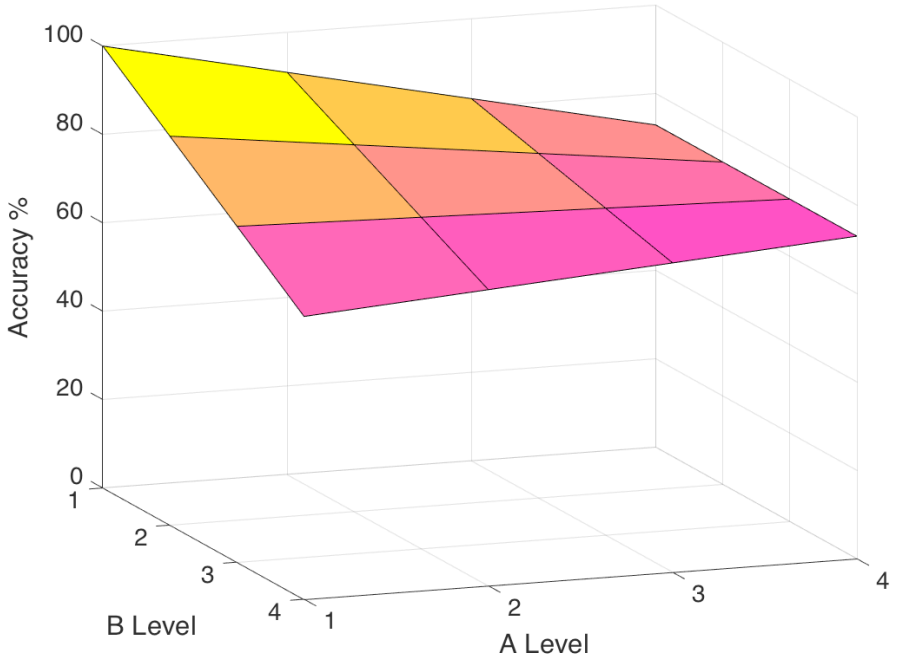

Figure 5.1 An illustration with simulated data of 3D regression models. In these models the dependent variable is Accuracy (%), with Task Difficulties as independent variables, showing: (a) Tasks A x B interaction term = zero; (b) a negative Tasks A x B interaction term; and (c) a positive Tasks A x B interaction term. The Task A and Task B slope terms in all figures are negative.

Subjects' Task A, Task B, and Tasks A x B regression coefficients for each of the five dual task variables served as the variables analysed in the current study Note that for brevity they are referred to by the same labels as the original variables:

- Subject's Accuracy;
- 1st Order Participant Accuracy Estimate - *Before* Dual Task Experience;
- 2nd Order Participant Accuracy Estimate - *Before* Dual Task Experience;
- 1st Order Participant Accuracy Estimate - *After* Dual Task Experience;
- 2nd Order Participant Accuracy Estimate - *After* Dual Task Experience;

but in the current study the variables refer to regression coefficients from the 3D models, and therefore now represent the degree which to the variable was affected by the rising difficulty levels of Tasks A, B, and A x B. The exception is the moderating variables of Subject's Metacognitive Accuracy for Task A, Task B, and Task A x B (described earlier) which are not regression coefficients but represent average values across the corresponding component of the dual task.

5.2.5 Research Questions

5.2.5.1 Primary Research Questions:

I. How did the increasing difficulty of Task A, Task B, and Tasks A x B affect subjects' accuracy on the dual task?

II. Did subjects expect another participant's accuracy on the dual task to be affected by the increasing difficulty of Task A, Task B, and Tasks A and B together?

III. Did the degree to which increasing task difficulties affected the subject's own accuracy predict their estimates for another participant, and was any relationship moderated by the subject's metacognitive accuracy?

5.2.5.2 Exploratory Research Question:

I. Did the subject's theory of mind ability, trait perspective-taking or self-esteem predict their estimates for another participant or further moderate Primary Research Question III?

5.2.6 Analysis Outline

One sample t-tests were performed for Primary Research Questions I and II. To test Primary Research Question III, a series of regression models were run separately for each component of the dual task: rising levels of (1) Task A, (2) Task B, and (3) Tasks A x B. In Step 1, the predictor variable was the degree to which Subject's Accuracy was affected by the rising levels of the relevant task; in Step 2, Subject's

Metacognitive Accuracy on the corresponding task was added as a moderating predictor; in Step 3, the interaction term between Subject's Accuracy and Subject's Metacognitive Accuracy was added to the model. For each component of the dual task, separate regression models were run for each of the four outcome variables, which were 1st and 2nd Order Participant Accuracy Estimates, both before and after experience of the dual task.

To investigate the Exploratory Research Question, subject's MASC % Accuracy, IRI-PT, and Rosenberg Self-esteem scores were added separately to the models described for Primary Research Question III, and stepwise regression analyses were performed, with sequential replacement, in the MASS package in R (Venables & Ripley, 2002). The model with best fit is reported where significant relationships were observed. Three outlying subjects, who were more than three standard deviations below the mean on the MASC % Accuracy, were removed for the relevant analyses.

5.3 Results

Descriptive statistics are shown in Figure 5.2 and Table 5.1.

Figure 5.2 a Subject's Accuracy

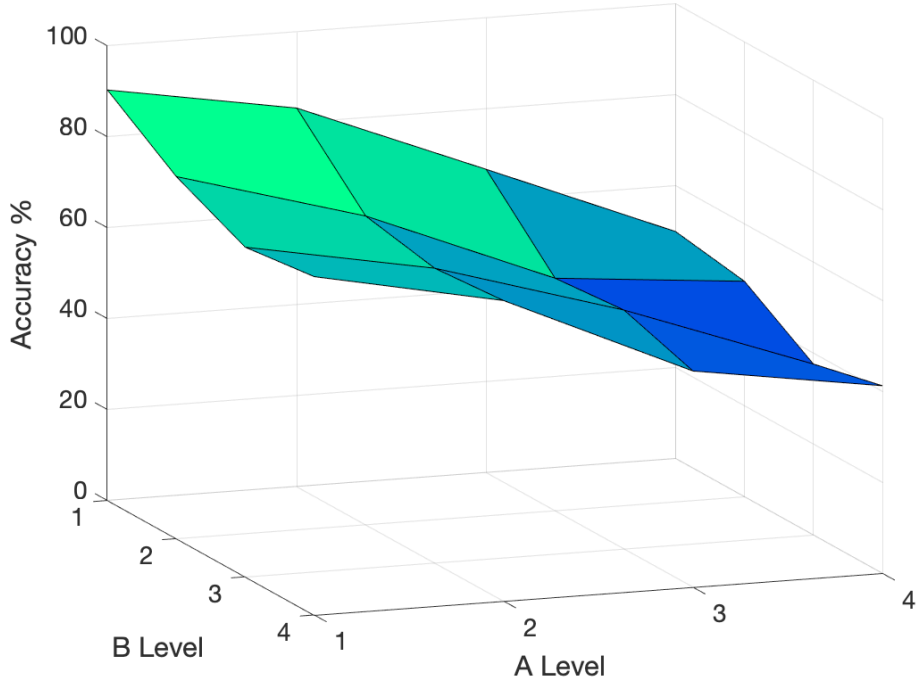

Figure 5.2 b 1st Order Participant Accuracy Estimate - *Before* Dual Task Experience

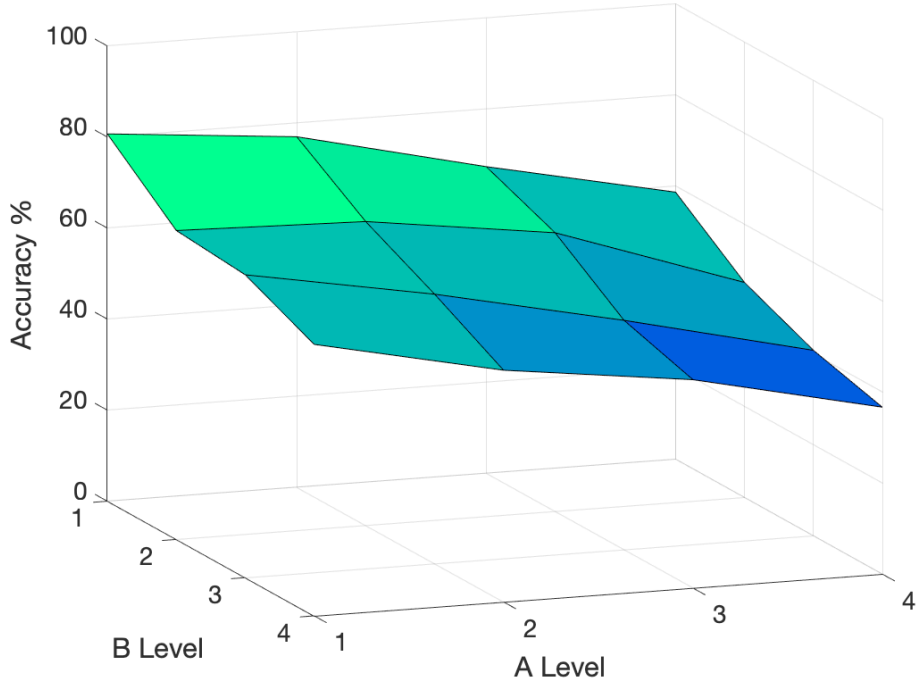

Figure 5.2 c 2nd Order Participant Accuracy Estimate - *Before* Dual Task Experience

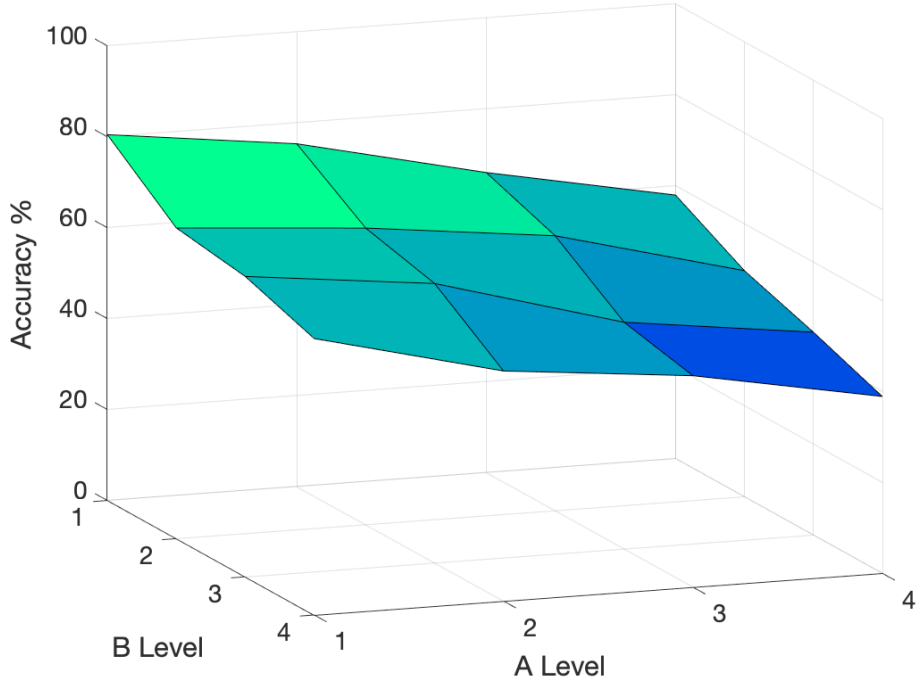

Figure 5.2 d 1st Order Participant Accuracy Estimate - *After* Dual Task Experience

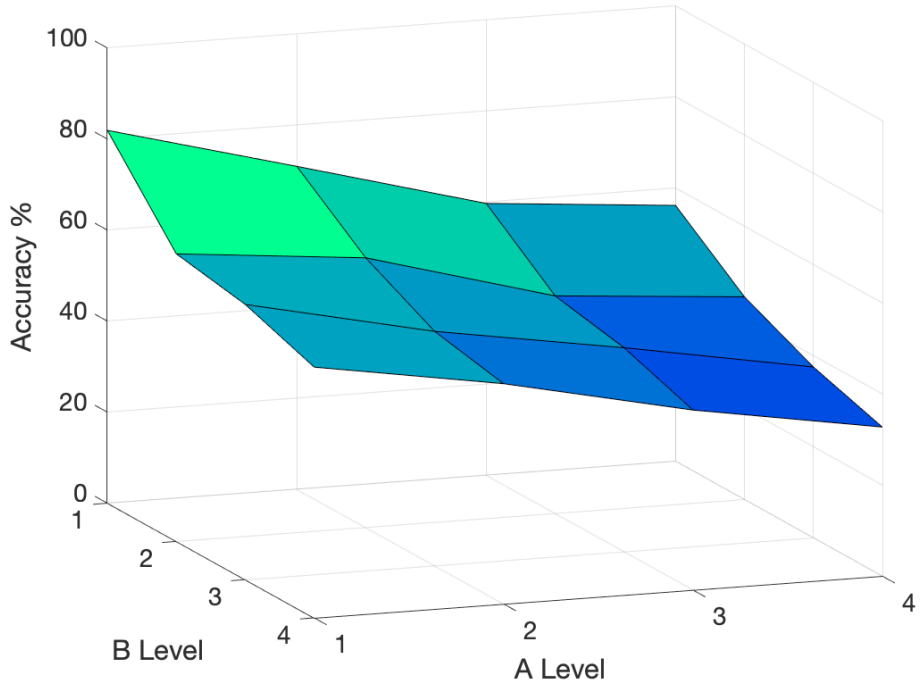

Figure 5.2 e 2nd Order Participant Accuracy Estimate - *After* Dual Task Experience

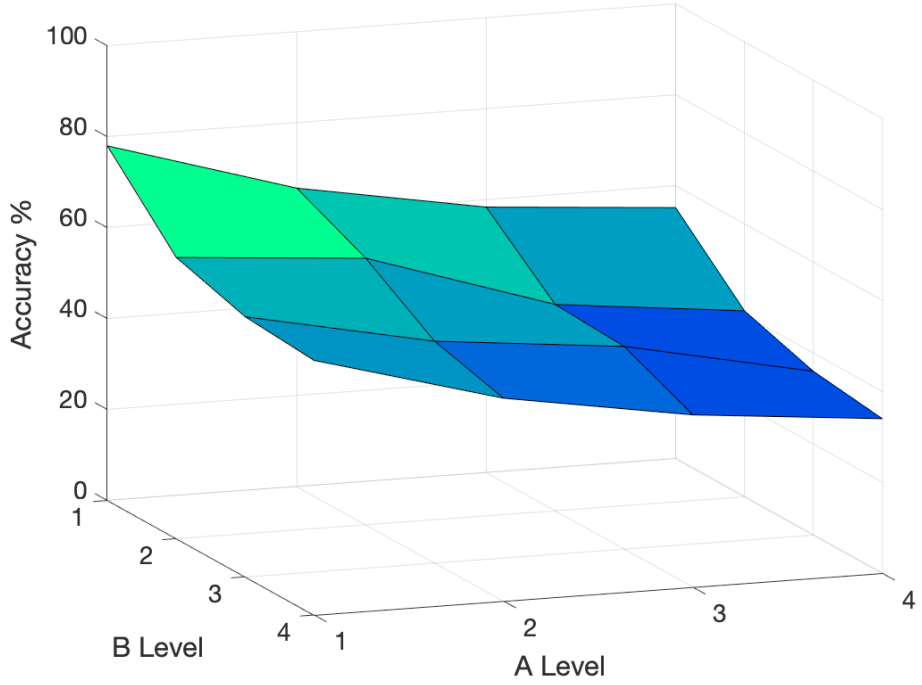

Figure 5.2 Average accuracy for each independent variable.

	Mean	SD	Range [min, max]
Task A Metacognitive Accuracy	0	15.01	[-18.02, 69.60]
Task B Metacognitive Accuracy	0	6.86	[-17.89, 20.70]
Tasks A x B Metacognitive Accuracy	0	8.35	[-12.47, 25.77]
Theory of Mind Ability (n = 64)	77.78	7.63	[60, 91]
Perspective-taking	24.28	4.51	[14, 35]
Self-esteem	21.48	5.66	[10, 40]

Table 5.1 Descriptive statistics for subjects' variables, N = 67.

5.3.1 Primary Research Question I:

How did the increasing difficulty of Task A, Task B, and Tasks A x B affect subjects' accuracy on the dual task? One-sample t-tests showed that subjects' accuracy on the dual task was significantly and negatively affected by the increasing difficulty of Task A (t (66) = -15.95, $p < .001$, $Mean$ = -13.51, 95% CI_{Mean} [-15.20, -11.82]) and Task B (t (66) = -8.62, $p < .001$, $Mean$ = -6.41, 95% CI_{Mean} [-7.89, -4.92]). There was a significant positive A x B interaction, (t (66) = 2.04, $p = .045$, $Mean$ = 0.55, 95% CI_{Mean} [0.01, 1.08]), indicating that accuracy was not as much impaired by the levels of Task A and B rising together as was predicted by the separate slopes of Task A and Task B (See Figure 5.2a.)

5.3.2 Primary Research Question II:

Did subjects expect another participant's accuracy on the dual task to be affected by the increasing difficulty of Task A, Task B, and Tasks A and B together? Subjects estimated that another's *accuracy* would decline as levels of Task A and Task B increased (Figure 5.2b). Subjects also thought that another participant's accuracy *estimates* would decrease as levels of Task A and Task B increased (Figure 5.2c). However, subjects did not estimate that Tasks A and B rising in difficulty together would impact on another participant's *accuracy,* or on their *estimate* thereof. Results for these one-sample t-tests are shown in Table 5.2. This pattern of results was observed for subjects' responses both before and after experiencing the dual task themselves (Figure 5.2d,e), paired-sample t-tests showed no significant difference in any accuracy estimate.

Taking Questions 1 and 2 together, these results show that, as a group, subjects: were affected by the rising levels of Task A and Task B; estimated that another participant's accuracy would also be affected; and estimated that the participant would

be aware of the effect on their accuracy. However, although subjects' accuracy was affected by Tasks A and B rising together, subjects did not estimate this effect for another participant's accuracy, or for that participant's estimate of their own accuracy, even after having specific experience of the dual task.

		Task A Difficulty	Task B Difficulty	Tasks A x B Difficulty
		M, *t*, *p*	*M*, *t*, *p*	*M*, *t*, *p*
Before	1st Order Participant Accuracy Estimate	-7.14, -5.96, < .001*	-7.09, -6.58, < .001*	-0.19, -0.47, .63
	2nd Order Participant Accuracy Estimate	-7.21, -5.75, < .001*	-6.88, -6.12, < .001*	0.04, 0.10, .92
After	1st Order Participant Accuracy Estimate	-7.94, -7.57, < .001*	-7.84, -6.88, < .001*	0.17, 0.49, .62
	2nd Order Participant Accuracy Estimate	-7.31, -7.13, < .001*	-6.95, -6.58, < .001*	0.08, 0.24, .81

Table 5.2 Results of one-sample t-tests for Primary Research Question II.
The variables are subjects' 1st and 2nd order estimates of participant's accuracy both before and after experience of the dual task.
***M* = Mean, *t* = t-test statistic, *p* = significance statistic, * < .05; all degrees of freedom = 66.**

5.3.3 Task A: Primary Research Question III

Did the degree to which increasing task difficulties affected the subject's own accuracy predict their estimates for another participant, and was any relationship moderated by the subject's metacognitive accuracy? Results are shown in Table 5.3. Before subjects experienced the dual task for themselves, the degree to which Subject's Accuracy on the dual task was affected by rising levels of Task A significantly and positively predicted the degree to which they estimated another participant's accuracy would be affected (1st Order Participant Accuracy). However there was no main effect of, or interaction with, Subject's Task A Metacognitive Accuracy (Steps 2 and 3, Table 5.3). There was no effect of Subject's Accuracy, Metacognitive Accuracy, or interaction thereof, on subject's estimation of another participant's estimation of how their accuracy would be affected (2nd Order Participant Accuracy).

After subjects experienced the dual task for themselves, the degree to which Subject's Accuracy on the dual task was affected by rising levels of Task A again significantly and positively predicted their estimate of the degree to which another participant's accuracy would be affected. There was no main effect of Subject's Metacognitive Accuracy, but there was now a significant interaction between Subject's Accuracy and Metacognitive Accuracy. For those with good metacognitive accuracy, there was no effect of Subject's Accuracy on 1st Order Participant Accuracy, B = 0.07, SD = 0.20, 95% CI [-0.34,0.48], $t = 0.33$, $p = .74$. Those with poor metacognitive accuracy showed a significant positive effect of Subject's Accuracy on 1st Order Participant Accuracy, B = 0.75, SD = 0.20, 95% CI [0.35,1.15], $t = 3.83$, $p = .001$*.

In contrast to before experience of the dual task, after such experience there was now a significant positive effect of Subject's Accuracy on subject's estimation of

another participant's estimation of how their accuracy would be affected (2nd Order Participant Accuracy). There was no effect of Metacognitive Accuracy and no interaction.

5.3.4 Task A: Exploratory Research Question

Did subject's theory of mind ability, trait perspective-taking or self-esteem predict their estimates for another participant or further moderate Primary Research Question III? After experience of the dual task, there was a significant interaction between Subject's Task A Metacognitive Accuracy and Subject's MASC % Accuracy on predicting the degree to which subjects estimated another participant's accuracy would be affected (1st Order Participant Accuracy – After Dual Task Experience), B = 0.03, SD = 0.01, t = 2.17, p = .03*. For those with poor metacognitive accuracy, there was no effect of subject's MASC % Accuracy, B = -0.15, SD = 0.34, t = -0.45, p = .66. In contrast, for those with good metacognitive accuracy, MASC % Accuracy significantly and negatively predicted subject's estimates of the degree to which another participant's accuracy would be affected by the rising levels of Task A, B = -0.97, SD = 0.47, t = -2.08, p = .046*. This means that, among those who had good metacognitive accuracy, those who had better theory of mind ability were more likely to estimate that another participant's accuracy would be affected by the increasing levels of Task A.

For the same outcome variable (1st Order Participant Accuracy – After Dual Task Experience), a three-way interaction of Subject's Accuracy x Metacognitive Accuracy x Self-esteem was also observed, B = 0.006, SD = 0.003, t = 2.08, p = .04*. For those with poor metacognitive accuracy and high self-esteem, Subject's Accuracy significantly and positively predicted the degree to which subjects estimated another participant's accuracy would be affected, B = 0.99, SD = 0.23, t = 4.23, p = .001*. No

significant relationship was observed for those with good metacognitive accuracy and high self-esteem, or for those with low self-esteem irrespective with either good or poor metacognitive accuracy.

5.3.5 Task B: Primary Research Question III

Results are shown in Table 5.4. Before subjects experienced the dual task for themselves, the degree to which Subject's Accuracy on the dual task was affected by rising levels of Task B, and Task B Metacognitive Accuracy, did not predict the degree to which they estimated another participant's accuracy would be affected, but there was a marginally significant (p = .05) interaction effect. For those with good metacognitive accuracy, there was no effect of Subject's Accuracy on 1st Order Participant Accuracy, B = -0.02, SD = 0.27, 95% CI [-0.57,0.53], t = -0.07, p = .94, but those with poor metacognitive accuracy showed a marginally significant and positive effect of Subject's Accuracy on 1st Order Participant Accuracy, B = 0.47, SD = 0.23, 95% CI [0.00,0.94], t = 2.04, p = .05. There were no significant predictors for 2nd Order Participant Accuracy, nor for 1st and 2nd Order Participant Accuracy after experiencing the dual task.

5.3.6 Task B: Exploratory Research Question

No significant predictors were observed.

5.3.7 Tasks A x B: Primary Research Question III

Results are shown in Table 5.5. There were no significant effects of the degree to which Subject's Accuracy on the dual task was affected by levels of Tasks A and B rising together, and Subject's Tasks A x B Metacognitive Accuracy, on any of the outcome variables. One marginally significant result to note is that, before subjects experienced the dual task for themselves, Subject's A x B Accuracy predicted their estimate of the degree to which another participant's accuracy would be affected (1st Order Participant Accuracy).

5.3.8 Tasks A x B: Exploratory Research Question

Before experience of the dual task, there was a significant interaction between the degree to which Subject's Accuracy on the dual task was affected by rising levels of Tasks A x B and Subject's MASC % Accuracy on predicting the subject's estimation of another participant's estimation of how their accuracy would be affected (2nd Order Participant Accuracy), B = 0.05, SD = 0.02, t = 2.24, p = .03*. For those with poor MASC % Accuracy, Subject's A x B Accuracy significantly and negatively predicted the subject's estimation of another participant's estimation of how their accuracy would be affected, B = -1.07, SD = 0.46, t = -2.32, p = .048*. This means that, for those with poor theory of mind ability, the less their own accuracy was actually affected, the more likely they were to estimate that another participant would estimate that their accuracy would be negatively affected by Tasks A and B rising together. No relationship was observed for those with good MASC % Accuracy.

After experience of the dual task, there was a significant three-way interaction between the degree to which Subject's Accuracy on the dual task was affected by rising levels of Tasks A x B, Subject's A x B Metacognitive Accuracy, and Subject's Perspective-taking on predicting the degree to which they estimated another participant's accuracy would be affected (1st Order Participant Accuracy), B = 0.01, SD = 0.01, t = 2.04, p = .046*. For those with good metacognitive accuracy and high perspective-taking tendencies, there was a marginally significant negative effect of Subject's A x B Accuracy, B = -0.58, SD = 0.30, t = -1.94, p = .08. This means that the less their own accuracy was actually affected, the more likely they were to estimate that another participant's accuracy would decrease as Tasks A and B rose together. No relationship was observed for those with poor metacognitive accuracy and high perspective-taking tendencies, or for those with low perspective-taking tendencies with either good or poor metacognitive accuracy.

For the same outcome variable (1st Order Participant Accuracy – After Dual Task Experience), a significant interaction between Subject's A x B Accuracy and Subject's Self-esteem was observed, B = 0.06, SD = 0.03, t = 2.08, p = .04. For those with high self-esteem, Subject's Accuracy marginally significantly and positively predicted subject's estimates of another participant's accuracy, B = 0.62, SD = 0.32, t = 1.97, p = .06. No relationship was observed for those with low self-esteem.

		Before Dual Task Experience B (SD) [95% CI] t, p		After Dual Task Experience B (SD) [95% CI] t, p	
	Predictors	1st Order Participant Accuracy Estimate	2nd Order Participant Accuracy Estimate	1st Order Participant Accuracy Estimate	2nd Order Participant Accuracy Estimate
Step 1	Subject's AC	0.47 (0.17) [0.15, 0.80]	0.21 (0.18) [-0.16, 0.57]	0.40 (0.15) [0.11, 0.69]	0.30 (0.15) [0.01, 0.59]
		2.87, .006*	1.14, .26	2.79, .007*	2.03, .046*
Step 2	Subject's AC	0.49 (0.16) [0.16, 0.82]	0.22 (0.18) [-0.15, 0.58]	0.42 (0.14) [0.13, 0.71]	0.30 (0.15) [0.01, 0.60]
		2.97, .004*	1.19, .24	2.89, .005*	2.07, .04*
	Subject's MCA	-0.10 (0.08) [-0.25, 0.05]	-0.07 (0.08 [-0.23, 0.10]	-0.09 (0.07) [-0.22, 0.04]	-0.04 (0.07) [-0.18, 0.09]
		-1.29, .20	-0.77, .44	-1.35, .18	-0.65, .52
Step 3	Subject's AC	0.49 (0.17) [0.16, 0.83]	0.20 (0.18) [-0.17, 0.57]	0.37 (0.14) [0.09, 0.64]	0.28 (0.10) [-0.01, 0.58]
		2.95, .004*	1.09, .28	2.69, .009*	1.92, .06
	Subject's MCA	-0.12 (0.11) [-0.34, 0.11]	-0.00 (0.13) [-0.25, 0.25]	0.13 (0.09) [-0.05, 0.32]	0.04 (0.10) [-0.16, 0.24]
		-1.01, .32	-0.02, .99	1.44, .15	0.43, .67
	Subject's AC *	-0.00 (0.01) [-0.02, 0.02]	0.01 (0.01) [-0.01, 0.03]	0.02 (0.01) [0.01, 0.04]	0.01 (0.01) [-0.01, 0.02]
	Subject's MCA	-0.21, .84	0.68, .50	3.26, .002*	1.18, .24

Table 5.3 Results of the regression models for Primary Research Question III.
Variables reflect the dependence on the difficulty of Task A. AC = Accuracy; MCA = Metacognitive Accuracy. *p < .05.

		Before Dual Task Experience B (SD) [95% CI] t, p		After Dual Task Experience B (SD) [95% CI] t, p	
	Predictors	1st Order Participant Accuracy Estimate	2nd Order Participant Accuracy Estimate	1st Order Participant Accuracy Estimate	2nd Order Participant Accuracy Estimate
Step 1	Subject's AC	0.26 (0.18) [-0.09, 0.62] 1.48, .14	-0.18 (0.19) [-0.56, 0.19] -0.99, .33	0.19 (0.19) [-0.19, 0.56] 0.99, .32	0.09 (0.18) [-0.26, 0.44] 0.50, .62
Step 2	Subject's AC	0.25 (0.18) [-0.10, 0.60] 1.41, .16	-0.20 (0.19) [-0.57, 0.17] -1.08, .29	0.18 (0.19) [-0.20, 0.56] 0.96, .34	0.09 (0.18) [-0.27, 0.45] 0.51, .61
	Subject's MCA	-0.19 (0.16) [-0.50, 0.12] -1.22, .23	-0.24 (0.16) [-0.56, 0.09] -1.45, 0.16	-0.07 (0.17) [-0.40, 0.27] -0.40, .69	0.03 (0.16) [-0.28, 0.35] 0.20, .84
Step 3	Subject's AC	0.22 (0.17) [-0.13, 0.57] 1.26, 0.21	-0.22 (0.19) [-0.59, 0.15] -1.18, .24	0.17 (0.19) [-0.21, 0.55] 0.88, .38	0.08 (0.18) [-0.28, 0.44] 0.44, .66
	Subject's MCA	0.05 (0.20) [-0.34, 0.45] 0.28, .78	-0.09 (0.21) [-0.51, 0.34] -0.40, 0.69	0.04 (0.22) [-0.40, 0.48] 0.18, .86	0.12 (0.20) [-0.29, 0.53] 0.60, .55
	Subject's AC *	0.05 (0.03) [-0.00, 0.10]	0.03 (0.03) [-0.02, 0.09]	0.02 (0.03) [-0.04, 0.08]	0.02 (0.03) [-0.04, 0.07]
	Subject's MCA	1.96, 0.05	1.12, .27	0.78, .44	0.71, .48

Table 5.4 Results of the regression models for Primary Research Question III.
Variables reflect the dependence on the difficulty of Task B. AC = Accuracy; MCA = Metacognitive Accuracy. *p < .05.

		Before Dual Task Experience B (SD) [95% CI] t, p		After Dual Task Experience B (SD) [95% CI] t, p	
	Predictors	1st Order Participant Accuracy Estimate	2nd Order Participant Accuracy Estimate	1st Order Participant Accuracy Estimate	2nd Order Participant Accuracy Estimate
Step 1	Subject's AC	0.32 (0.18) [-0.04, 0.67] 1.76, .08	0.07 (0.20) [-0.31, 0.48] 0.43, .67	0.19 (0.16) [-0.13, 0.51] 1.17, .25	0.06 (0.15) [-0.23, 0.35] 0.42, .68
Step 2	Subject's AC	0.32 (0.18) [-0.04, 0.68] 1.78, .08	0.09 (0.20) [-0.31, 0.49] 0.46, .65	0.20 (0.16) [-0.12, 0.51] 1.22, .23	0.07 (015) [-0.23, 0.36] 0.45, .66
	Subject's MCA	0.04 (0.05) [-0.06, 0.13] 0.78, .44	0.04 (0.05) [-0.06, 0.15] 0.78, .44	0.06 (0.04) [-0.03, 0.14] 1.40, .17	0.04 (0.04) [-0.04, 0.11] 0.98, .33
Step 3	Subject's AC	0.35 (0.19) [-0.03, 0.72] 1.83, .07	0.10 (0.21) [-0.31, 0.52] 0.50, .62	0.26 (0.17) [-0.07, 0.59] 1.57, .12	0.06 (0.15) [-0.25, 0.36] 0.37, .71
	Subject's MCA	0.03 (0.05) [-0.07, 0.13] 0.53, .60	0.04 (0.06) [-0.08, 0.15] 0.62, .54	0.03 (0.05) [-0.06, 0.13] 0.77, .44	0.04 (0.04) [-0.04, 0.12] 0.98, .33
	Subject's AC *	(0.03) [-0.05, 0.08]	0.01 (0.03) [-0.06, 0.07]	0.04 (0.03) [-0.02, 0.09]	-0.00 (0.02) [-0.05, 0.04]
	Subject's MCA	0.48, .64	0.22, 0.83	1.36, .18	-0.20, .84

Table 5.5 Results of the regression models for Primary Research Question III.
Variables reflect the dependence on the difficulty of Tasks A and B together. AC = Accuracy; MCA = Metacognitive Accuracy.
***p < .05.**

5.4 General Discussion

This study attempted to examine the role of the self and metacognition in modelling another's memory. We therefore designed a dual task with levels of increasing difficulty and asked subjects to predict another's accuracy and estimate of such, both before and after having specific experience of the dual task themselves. They key questions were: whether subjects' own accuracy predicted their estimates for another; whether their metacognitive accuracy moderated the extent to which they did so; and was this further moderated by subject's theory of mind ability or trait self-esteem or perspective-taking tendencies. Overall, the findings do not show a consistent and clear pattern but, as an exploratory first study, offer some partial insights in response to the questions posed.

Across the whole sample, the findings show that subjects were affected by the increasing task difficulty and predicted that another person's accuracy would also be affected and that they would be aware of this. This accords with research on working memory capacity (Baddeley, 2010; Conway, Kane, & Engle, 2003), and with findings that show that dual tasks do affect performance when, as in the current study, both tasks call on the same memory resources as opposed to distinct resources (e.g. verbal vs. visuospatial; Cocchini, Logie, Della Sala, MacPherson, & Baddeley, 2002). However, subjects were insensitive to the small positive effect on accuracy of both tasks rising together which, although present in this group of subjects, may not be found in a larger group of subjects.

In the absence of any information about the other participant, we explored whether the subject's own accuracy predicted their estimates for them. For the Letter Span task, own performance did predict the subject's estimates of another's accuracy, but only after experience of the dual task did it affect 2nd order predictions of what the

other participant would predict for themselves. No such relationships were observed for the N-back task, except for a marginal interaction with metacognitive accuracy (discussed below) for the two tasks rising together. This indicates that predictions of other's memory performance may be task-specific. It is possible that subjects were more familiar with the letter span task as it is similar to everyday tasks of recalling spellings or phone numbers, whereas the N-back task is a novel experience. Magnussen et al (2006) have found that whether people's beliefs about memory concur with results from the research literature depends on the type of memory in question.

We next explored the possible moderating effects of meta-cognitive accuracy on the relationship between self-accuracy and estimates of another's accuracy. For the Letter Span task, meta-cognitive accuracy only moderated the effect of the self's accuracy on subject's predictions of another's accuracy after experiencing the dual task. In contrast, for the N-back task, this effect was observed before experiencing the dual task. In both cases, those with poor meta-cognitive accuracy showed a positive relationship between their own accuracy and their predictions of another's. This is not an intuitive finding, as it indicates a relationship between estimates of other's accuracy and one's actual accuracy despite an inaccurate understanding of one's own accuracy. Possibly this is due to a bias in one's metacognition. For instance, Magnussen et al (2006) found that 70% of people rated themselves as good or very good at judging the reliability of their own memory, despite their research being motivated by considerable evidence of low correlations between confidence and accuracy in eyewitness testimonies. Alternatively, it may be that estimates implicitly comprised both a judgement of the self and of the task difficulty. When judging one's own performance the self component may have additionally affected the response, but when judging another's performance only the task difficulty component was utilised. If the self's own

accuracy in fact corresponded to the task difficulty then this would explain how, despite having poor awareness of it, one's own accuracy could predict estimates of another's.

Indeed, to explore the effects of possible biases we investigated moderating effects of trait self-esteem and perspective-taking tendencies. Again, no consistent pattern appeared and these findings are purely exploratory. On the Letter Span task, after experience of the dual task, those with high self-esteem but poor metacognitive accuracy had a positive relationship between own accuracy and estimates of another's. Across both tasks together, those with high self-esteem also showed the same positive relationship. Indeed, the Rosenberg Self-esteem scale includes two items that measure whether the self's worth and abilities is equal to (not better than) that of others (Gray-Little et al. 1997). The only finding for trait perspective-taking showed that, after experiencing the dual task, those with good metacognitive accuracy and high perspective-taking tendencies had a negative relationship between one's own accuracy and one's estimate of another's. This implies that they accurately represented their own performance, but having experienced the task they estimated that others would perform worse than they did. This possibly suggests an increased self-other distinction, not using the self as a model for others.

The Mind-space framework suggests a way to study the relationship between metacognition of one's own and other's cognitive processes and mental state representation. However, in this exploratory study we only examined whether performance on an established theory of mind task (Dziobek et al., 2006) moderated the modelling of another's memory. Once again, no clear, consistent pattern was found from which to draw firm conclusions. In the Letter Span task, only among those with good metacognitive accuracy, those with better theory of mind ability were more likely to estimate that another participant's accuracy would be affected by the increasing

difficulty levels. For the two tasks together, among those with poor theory of mind ability, the less their own accuracy was actually affected, the more likely they were to estimate that another participant would think their accuracy would drop with increasing task difficulty. These findings offer some small support for the relationship between representing other's mental states and cognitive processes. As discussed in Section 2.6, the recipient design literature shows that people adjust their communicative behaviour based on their beliefs about the age of their partner, and the extent to which they do so is related to empathy (Newman-Norlund et al., 2009) and social experience (Stolk, Hunnius, Bekkering, & Toni, 2013). Although not well supported by the current study, we suggest that such recipient design findings reflect the representation of others' cognitive processes in addition to their mental state representations; that these two aspects are inherently linked. Future studies, designed more specifically to measure the mapping between location on cognitive Mind-space dimensions and mental state representations, are necessary to directly address this claim.

It is notable that there was no systematic effect of experience on estimates of other's accuracy. This suggests that prior beliefs about memory had a greater effect on estimations than experience. Correspondingly, Kornell and colleagues (Kornell, Rhodes, Castel, & Tauber, 2011) found that beliefs about memory and metamemory predictions had low correspondence to actual memory task performance; Irak and Çapan (2018) demonstrated that the relationship between metacognitive confidence and actual memory performance was mediated by beliefs about memory; and, as mentioned earlier, Magnussen et al (2006) found that beliefs accurately reflected memory performance only on certain tasks. Overall, the current study supports the idea that concepts of others' memory are also more influenced by pre-existing beliefs than recent experience.

There are several significant limitations to the current study. An initial limitation is the possibility that some decrement in performance on the dual task was in fact due to competition for sensory input, and not the dual task demands per se (Cocchini et al., 2002). Although the numbers and letters were aligned so that it was not necessary to saccade to see them, other studies have better avoided this confound by using preloading memory tasks, for instance holding information in mind while performing another task (Cocchini et al., 2002). More importantly, this study acts only as an initial exploratory investigation into the relationship between representing cognitive processes and mental states, to bridge the metacognitive and theory of mind literatures. The design was inadequate to fully address this, and future studies are likely to require significant redesigning.

A particular limitation, to the study of social representations in general, is the lack of any ground truth value from which to compute the 'accuracy' of other people. One avenue for future research would be to use the well-validated cognitive task scores from Intelligence Quotient tests (Wechsler, 2011) to represent the population values for average cognitive performance. Indeed, such IQ tests can provide percentiles and age-specific estimates which allows for the investigation of representations of specific categories of 'other', e.g. young children vs. older adults. This line of research would have important applications, for instance in reducing the use of elderspeak to older adults (Williams, Kemper, & Hummert, 1995), or improving teaching methods through better models of children's mental processing and their epistemic states.

The strengths of the current study lie in it being the first attempt to investigate a non-social, cognitive dimension in Mind-space. Moreover, the theoretical point - that the Mind-space framework can inform *when* the self is a good model for others and *why* metacognition of one's own mind may affect representations of the self's and others'

mental states – can generate new studies that can go beyond existing explanations of egocentric bias or simulation theory. The inadequacy of the current study serves as a guide to future Mind-space studies to better match the study design to the theoretical intent.

5.5 References

Baddeley, A. (2010). Working memory. *Current Biology*, *20*(4), R136-R140.

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

Cocchini, G., Logie, R. H., Della Sala, S., MacPherson, S. E., & Baddeley, A. D. (2002). Concurrent performance of two memory tasks: Evidence for domain-specific working memory systems. *Memory and Cognition*, *30*(7), 1086–1095. https://doi.org/10.3758/BF03194326

Conway, A. R. A., Kane, M. J., & Engle, R. W. (2003). Working memory capacity and its relation to general intelligence. *Trends in Cognitive Sciences*, *7*(12), 547–552. https://doi.org/10.1016/j.tics.2003.10.005

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

Dunlosky, J., & Tauber, S.U. (2015). *The Oxford Handbook of Metamemory*. Oxford: Oxford Univeristy Press

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

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

Flavell, J. H. (2000). Development of children's knowledge about the mental world. *International Journal of Behavioral Development*, *24*(1), 15–23. https://doi.org/10.1080/016502500383421

Frith, U., & Happé, F. (1999). Theory of mind and self-consciousness: What is it like to be autistic? *Mind & Language*, *14*(1), 1–22. https://doi.org/10.1111/1468-0017.00100

Grainger, C., Williams, D. M., & Lind, S. E. (2014). Metacognition, metamemory, and mindreading in high-functioning adults with autism spectrum disorder. *Journal of Abnormal Psychology*, *123*(3), 650–659. https://doi.org/10.1037/a0036531

Gray-Little, B., Williams, V. S. L., & Hancock, T. D. (1997). An Item Response Theory Analysis of the Rosenberg Self-Esteem Scale. *Personality and Social Psychology Bulletin*, *23*(5), 443–451.

Irak, M., & Çapan, D. (2018). Beliefs about Memory as a Mediator of Relations between Metacognitive Beliefs and Actual Memory Performance. *Journal of General Psychology*, *145*(1), 21–44. https://doi.org/10.1080/00221309.2017.1411682

Kornell, N., Rhodes, M. G., Castel, A. D., & Tauber, S. K. (2011). The ease-of-processing heuristic and the stability bias: Dissociating memory, memory beliefs,

and memory judgments. *Psychological Science*, *22*(6), 787–794. https://doi.org/10.1177/0956797611407929

Leslie, A. M. (1987). Pretense and Representation: The Origins of "Theory of Mind". *Psychological Review*, *94*(4), 412–426. https://doi.org/10.1037/0033-295X.94.4.412

Magnussen, S., Andersson, J., Cornoldi, C., De Beni, R., Endestad, T., Goodman, G. S., … Zimmer, H. (2006). What people believe about memory. *Memory*, *14*(5), 595–613. https://doi.org/10.1080/09658210600646716

Nelson, T. O., Narens, L., & Dunlosky, J. (2004). A Revised Methodology for Research on Metamemory: Pre-judgment Recall And Monitoring (PRAM). *Psychological Methods*, *9*(1), 53–69. https://doi.org/10.1037/1082-989X.9.1.53

Newman-Norlund, S. E., Noordzij, M. L., Newman-Norlund, R. D., Volman, I. A. C., Ruiter, J. P. de, Hagoort, P., & Toni, I. (2009). Recipient design in tacit communication. *Cognition*, *111*(1), 46–54. https://doi.org/10.1016/j.cognition.2008.12.004

Perner, J. and Kuhlberger, A. (2005). Mental Simulation: Royal road to other minds? In Other Minds (Malle, B.F. and Hodges, S.D., eds), New York: Guilford Press

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

Shea, N., Boldt, A., Bang, D., Yeung, N., Heyes, C., & Frith, C. D. (2014). Supra-personal cognitive control and metacognition. *Trends in Cognitive Sciences*, *18*(4),

186–193. https://doi.org/10.1016/j.tics.2014.01.006

Stolk, A., Hunnius, S., Bekkering, H., & Toni, I. (2013). Early Social Experience Predicts Referential Communicative Adjustments in Five-Year-Old Children. *PLoS ONE*, *8*(8), e72667. https://doi.org/10.1371/journal.pone.0072667

Venables, W. N. & Ripley, B. D. (2002). *Modern Applied Statistics with S. Fourth Edition*. Springer, New York. ISBN 0-387-95457-0

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

Williams, D. (2010). Theory of own mind in autism: Evidence of a specific deficit in self-awareness? *Autism : The International Journal of Research and Practice*, *14*(5), 474–494. https://doi.org/10.1177/1362361310366314

Williams, K., Kemper, S., & Hummert, M. L. (1995). Practice Concepts Improving Nursing Home Communication : An Intervention to Reduce Elderspeak. *The Gerontologist*, *43*(2), 242–247. https://doi.org/10.1093/geront/43.2.242

6. Submentalizing or Mentalizing in a Level 1 Perspective-Taking Task: A Cloak and Goggles Test

This chapter is presented as a published article and is an exact copy of the following journal publication:

Conway, J.R., Lee, D., Ojaghi, M., Catmur, C. & Bird, G. (2017). Submentalizing or Mentalizing in a Level 1 Perspective-Taking Task: A Cloak and Goggles Test. *Journal of Experimental Psychology: Human Perception and Performance*, *43*(3), 454-465. https://doi.org/ 10.1037/xhp0000319

Corresponding author:

J.R. Conway, Social, Genetic and Developmental Psychiatry Centre, Institute of Psychiatry, Psychology and Neuroscience, King's College London, London SE5 8AF, United Kingdom.

jane_rebecca.conway@kcl.ac.uk

Videos related to the methods can be viewed at: https://osf.io/jas4n/

Journal of Experimental Psychology:
Human Perception and Performance
2017, Vol. 43, No. 3, 454–465

0096-1523/17/$12.00 http://dx.doi.org/10.1037/xhp0000319

Submentalizing or Mentalizing in a Level 1 Perspective-Taking Task: A Cloak and Goggles Test

Jane R. Conway, Danna Lee, Mobin Ojaghi, and Caroline Catmur
King's College London

Geoffrey Bird
King's College London and University College London

It has been proposed that humans possess an automatic system to represent mental states ('implicit mentalizing'). The existence of an implicit mentalizing system has generated considerable debate however, centered on the ability of various experimental paradigms to demonstrate unambiguously such mentalizing. Evidence for implicit mentalizing has previously been provided by the 'dot perspective task,' where participants are slower to verify the number of dots they can see when an avatar can see a different number of dots. However, recent evidence challenged a mentalizing interpretation of this effect by showing it was unaltered when the avatar was replaced with an inanimate arrow stimulus. Here we present an extension of the dot perspective task using an invisibility cloaking device to render the dots invisible on certain trials. This paradigm is capable of providing unambiguous evidence of automatic mentalizing, but no such evidence was found. Two further well-powered experiments used opaque and transparent goggles to manipulate visibility but found no evidence of automatic mentalizing, nor of individual differences in empathy or perspective-taking predicting performance, contradicting previous studies using the same design. The results cast doubt on the existence of an implicit mentalizing system, suggesting that previous effects were due to domain-general processes.

Public Significance Statement
The ability to represent in one's own mind what other people see, think, or believe is important for social interactions and relationships. There is wide agreement that this 'mentalizing' ability depends on a late developing, slow and effortful system, but much debate on whether humans also possess a fast and automatic mentalizing system. The present studies tested whether participants automatically represented what an onscreen human avatar could see. Objects' visibility was manipulated by using either a set of telescopes or goggles. One of each set allowed objects to be seen, and the other did not. Participant response times were predicted to be faster when what they saw corresponded to what the avatar saw, and slower when there was a difference. However, this did not occur, providing no evidence for an automatic mentalizing system, suggesting rather that representing others' mental states is effortful not automatic.

Keywords: domain-general processing, implicit mentalizing, submentalizing, theory of mind, visual perspective taking

Supplemental materials: http://dx.doi.org/10.1037/xhp0000319.supp

This article was published Online First November 28, 2016.

Jane R. Conway, Danna Lee, and Mobin Ojaghi, MRC Social, Genetic, & Developmental Psychiatry Centre, Institute of Psychiatry, Psychology, & Neuroscience, King's College London; Caroline Catmur, Department of Psychology, Institute of Psychiatry, Psychology, & Neuroscience, King's College London; Geoffrey Bird, MRC Social, Genetic, & Developmental Psychiatry Centre, Institute of Psychiatry, Psychology, & Neuroscience, King's College London, and Institute of Cognitive Neuroscience, University College London.

This work was supported by an Economic and Social Research Council studentship [Ref: 1413340] awarded to Jane R. Conway. Jane R. Conway, Caroline Catmur, and Geoffrey Bird contributed to the study concept and design. Jane R. Conway and Mobin Ojaghi built the cloaking device, with additional help provided by J. Choi, T. Meany, B. Xia, and R. Howells. Data collection was performed by Jane R. Conway and Danna Lee. Data analysis and interpretation were performed by Jane R. Conway under the supervision of Geoffrey Bird and Caroline Catmur. Jane R. Conway drafted the manuscript, and Danna Lee, Caroline Catmur, and Geoffrey Bird provided critical revisions. All authors approved the final version of the manuscript for submission.

We are exceedingly grateful to Ian Apperly and Tiziano Furlanetto for generously sharing their experimental materials and for helpful discussions. We are also grateful to Cecilia Heyes for discussion of methods to assess implicit or automatic mentalizing.

Correspondence concerning this article should be addressed to Jane R. Conway, MRC Social, Genetic, & Developmental Psychiatry Centre, Institute of Psychiatry, Psychology, & Neuroscience, King's College London, London, UK. E-mail: jane_rebecca.conway@kcl.ac.uk

Mentalizing (also known as 'theory of mind') refers to the ascription of mental states, such as beliefs and intentions, to oneself and others. Mentalizing plays a crucial role in social interactions, particularly when seeking to predict, understand, or explain another's behavior. Although the existence of a late-developing, cognitively demanding, ability to represent mental states in human adults and older children is almost universally accepted, there has been considerable debate regarding the existence of an earlier, more automatic and efficient route by which infants and nonhuman animals may represent beliefs, or even 'belief-like' states (Apperly & Butterfill, 2009; Butterfill & Apperly, 2013). This debate has largely been methodological in nature, with several authors claiming evidence for an automatic mentalizing system (sometimes described as an 'implicit mentalizing' system—Kovács, Téglás, & Endress, 2010; Qureshi, Apperly, & Samson, 2010; Samson, Apperly, Braithwaite, Andrews, & Bodley Scott, 2010; Senju, Southgate, Snape, Leonard, & Csibra, 2011), whereas others have provided alternative explanations for the effects claimed to support the existence of such a system (Cole, Atkinson, Le, & Smith, 2016; Heyes, 2014a; Phillips et al., 2015; Santiesteban, Catmur, Hopkins, Bird, & Heyes, 2014).

In adults, the 'dot perspective task' has provided some of the strongest evidence that mentalizing can occur automatically (Qureshi et al., 2010; Samson et al., 2010). Participants are presented with an image of a blue room with red dots on the wall. In the center a human avatar faces toward the right or left wall. Participants are asked to verify whether a given number cue matches the number of red dots they themselves can see on the walls of the room. Importantly, they are instructed to ignore the avatar and respond based on their own visual perspective. On *consistent* trials the number of dots the participant and avatar can see is the same; on *inconsistent* trials the participant and avatar see a different number of dots (because some of the dots are positioned behind the avatar). Despite being told to ignore the avatar, participants respond faster on consistent trials than on inconsistent trials. This 'consistency effect' has been interpreted as evidence for automatic mentalizing: that the avatar's visual perspective (i.e., mental state) is automatically processed in addition to the participant's own. It is suggested that on inconsistent trials, resolution of the conflict between the participant's and avatar's visual perspectives extends response times.

A limitation of the original dot perspective task's ability to provide evidence of automatic mentalizing is that it did not include a control condition that could test an alternative 'submentalizing' hypothesis (Heyes, 2014a). If mentalizing causes the consistency effect, then the effect should not be observed when the central stimulus is not an appropriate target for the attribution of mental states. However, a recent paper (Santiesteban et al., 2014) demonstrated that the same consistency effect is observed when the central stimulus is an arrow rather than an avatar. These data raise the question of whether the automatic process generating the consistency effect involves mentalizing—specifically, representation of what others can see—or a domain-general nonmentalistic process where, for example, the eyes/nose of the avatar and the point of the arrow act as directional cues that automatically orientate participants' attention to a subset of the dots, slowing responding on inconsistent trials (Catmur, Santiesteban, Conway, Heyes, & Bird, 2016; Santiesteban et al., 2014). A new experimental manipulation is therefore required to find positive evidence of automatic mentalizing in the avatar condition.

Heyes (1998, 2014b, 2015) proposed such a method, known as the 'goggles test,' that has provided the strictest test of mentalizing to date. The goggles test is the most refined of a general class of methods to identify mentalizing which make use of an opaque barrier to determine the ability to represent what another perceives. Barrier methods compare behavior in two situations: when in the presence of another agent with full visual access to the environment, and when in the presence of an agent whose view of the environment is blocked by an opaque barrier. In the goggles version of the test, participants first learn a conditional discrimination between two colored goggles, one of which affords seeing and the other not. Participants learn the affordances of the goggles through their own experience with them. A transfer test then follows where the goggles are placed on another individual and participants have to extrapolate from their own experience to infer what can be seen through each pair of goggles.

It could be argued, however, that successful performance on the goggles test does not provide an unequivocal demonstration of mentalizing. In common with other barrier methods, if a participant has repeated experience of opaque barriers, they may learn that when barriers are placed between an object and another individual then that individual does not interact with the object. This experience may allow them to act *as if* they realize that the individual does not see the object, and that therefore the individual does not know it is there, but does not require the participant to represent the other's mental state (Penn & Povinelli, 2007). One therefore needs to extend the goggles logic so that the participant encounters two situations in which an agent views a scene through transparent barriers, so that past experience suggests that the barrier affords seeing, but where one of the barriers renders a specific object invisible. This situation has been impossible to instantiate until now, but recent advances in the development of 'cloaking devices' to render objects invisible (Choi & Howell, 2014) make it possible.

Therefore, in Experiment 1, rather than giving participants experience of transparent and opaque goggles, we used two 'telescopes' within a cloaking device. The lenses in both telescopes were transparent, however because of their respective focal lengths it was possible to manipulate an object's visibility. An object placed at a specific distance from the focal point of the 'invisible telescope' was invisible, even though other objects not placed at that point were visible. All objects were visible when viewed through the 'visible telescope' because its lens had a different focal length. This cloaking device allowed us to maintain transparency across conditions while manipulating the visibility of a specific object: in this case, the red dots in the dot perspective task. The use of such a novel apparatus within the dot perspective task means that previous experience of how people interact with objects placed behind an opaque barrier cannot explain any modulation of the consistency effect by the visibility of the dots, and allows the inferred mental state of the avatar to be manipulated while holding all other task and stimulus features constant. The use of two telescopes, both with transparent lenses, one of which renders certain objects invisible, allows for precise manipulation of specific mental state content (i.e., what is seen). In addition, and unlike the goggles manipulation, the fact that participants do not have previous experience of transparent materials able to render

specific objects invisible means that any nonmentalistic explanation of their potential impact on the consistency effect based on prior learning becomes untenable.

Problems with prior experience of opacity notwithstanding, two recent studies used variants of the goggles test to determine whether automatic mentalizing underpins performance in the avatar condition of the dot perspective task. Cole et al. (2016) inserted either a transparent barrier or an opaque barrier in front of the avatar stimulus. This design provides a test of the automatic mentalizing hypothesis as if the consistency effect is attributable to the representation of the avatar's visual perspective, then the consistency effect should be modulated by anything that modulates the avatar's visual perspective (such as the opacity of the barrier). Counter to the hypothesis that automatic mentalizing underlies performance on the standard version of the task, participants demonstrated an equivalent consistency effect for both types of barrier.

Furlanetto, Becchio, Samson, and Apperly (2016) also implemented a variant of the goggles procedure in which they instructed participants as to the properties of two pairs of goggles, one of which was transparent and one opaque, before allowing them to experience the difference themselves. They then administered the standard version of the dot perspective task with the addition of conditions in which the avatar wore either the opaque or transparent goggles. Participants demonstrated a consistency effect when the avatar wore transparent goggles, but not when the avatar wore opaque goggles; results which are consistent with an automatic mentalizing interpretation.

The contrasting results between the Cole and Furlanetto studies can potentially be explained by a crucial methodological difference relating to the judgments participants were required to make during the dot perspective task. It has long been acknowledged that the 'acid test' of automatic mentalizing in this task occurs when participants are required to verify whether the number cue matches the number of dots visible from their own perspective only (as used by Cole et al., 2016). Here, any effect of the avatar meets a strict definition of automatic in which even though participants are never required to judge the number of dots visible from the avatar's perspective, and doing so hinders performance of the instructed task, their performance is nevertheless influenced by the avatar.

Other variants of the dot perspective task have required participants to verify the number cue from both their own and the avatar's perspective (as used by Furlanetto et al., 2016). The requirement to adopt both perspectives significantly weakens the claim for automaticity, as participants may experience task carry-over effects on own-perspective trials from avatar-perspective trials (Samson et al., 2010, p. 1259; Santiesteban et al., 2014, p. 934; Schurz et al., 2015, p. 387). Such an effect would be automatic in the sense that adoption of the avatar's perspective on own-perspective trials is task-irrelevant and interferes with performance, but the automaticity would be an artifact of the testing situation rather than a general feature of human cognition. Thus, it is possible that Furlanetto et al. have shown that a carry-over effect of explicit, nonautomatic mentalizing on the avatar-perspective trials modulates the consistency effect on self-perspective trials; however, *what* process is being modulated cannot be determined by their design: it could be either automatic mentalizing or a domain-general process.

Experiments 2 and 3 were designed to investigate this task carry-over explanation of the Furlanetto et al. (2016) result. Experiment 2 repeated the Furlanetto study using their exact design, stimuli, and procedure but with one key variation: participants were asked to respond from their own perspective only and never from the avatar's perspective. Experiment 3 was a replication of Furlanetto et al. (2016), in which participants responded from both their own and the avatar's perspective. Comparison of the results of Experiment 2 and Experiment 3 allows a task carry-over effect to be identified if it is present.

In sum, the current experiments utilized two different visibility manipulations embedded in the dot perspective task to determine whether the consistency effect is modulated by the avatar's inferred mental state. In contrast to the Cole and Furlanetto studies described above, in Experiment 1 participants were not instructed about the properties of the telescopes, instead they discovered their properties through self-discovery only (as per Heyes, 2014b). In addition, in Experiments 1 and 2 only own-perspective trials were used to limit the potential for task carry-over effects to explain the results. Experiment 3 included both self- and other-perspective trials to determine whether the process underpinning the consistency effect is modulated by a task carry-over effect of explicit, nonautomatic, mentalizing.

Experiment 1

Experiment 1 implemented a variant on the dot perspective task designed such that, should evidence of mentalizing be observed, this evidence could not be explained by submentalizing factors related to domain-general processes or task carry-over effects. This aim was achieved through the use of two clear glass 'telescopes' and the addition of an arrow stimulus as used by Santiesteban et al. (2014).

Participants were given real-life experience of the two telescopes, one visible and one invisible, in a blue room with red dots on the wall. Participants could see the red dots through the visible telescope, but not through the invisible telescope. Participants then completed the dot perspective task with the telescopes inserted in front of the avatar and arrow stimuli. If participants represent what the avatar can see, one would expect a consistency effect when the avatar is looking through the visible telescope because on consistent trials there is no conflict between the participant's and avatar's perspectives, but on inconsistent trials responding should be slowed due to the conflict in perspectives. However, a consistency effect would not be expected with the invisible telescope because even when the number of dots visible to the participant equals the number of dots in front of the avatar ('consistent trials'), the avatar cannot see the red dots through the invisible telescope and therefore the participant's and the avatar's perspectives are always in conflict. In effect, the use of the invisible telescope means that all trials are inconsistent, and therefore that response times (RTs) on 'consistent' and inconsistent trials should be equivalent. As the arrow is not an appropriate target for the attribution of mental states, no consistency effect should be observed with this stimulus, regardless of telescope type.

In contrast, if the consistency effect in the dot perspective task is a result of nonmentalistic domain-general processes, such as the directionality of the stimulus, then one would expect to observe a consistency effect for both the visible and invisible telescope in

both the avatar and arrow conditions, providing that the addition of the telescope stimulus does not impact on the relevant cue characteristic (such as the directionality of either stimulus).

Method

Participants. Forty-nine healthy adults volunteered to take part in this experiment in return for a small monetary sum. Data from six participants were excluded from the analysis, 3 because of a technical fault and a further 3 because of being outliers with respect to accuracy (error rate >25%). The remaining 43 participants (37 female) were aged between 17 and 48 ($M = 25.72$, $SD = 7.57$). The data-stopping rule and sample size were determined prior to data collection and were based on previous research. The target sample size was three times ($n = 48$) the size of the original dot perspective task study (Samson et al., 2010; $n = 16$).

Stimuli and apparatus.

Cloaking device. A real-life replica of the blue room from the computer stimuli of the dot perspective task was built. This room measured 275 mm high by 370 mm wide and was situated on an adjustable stand so participants could place their head inside the room while standing. A telescope mount was placed in the center of the room, 150 mm from its back wall. In the center of the back wall there was a porthole of 45 mm diameter where acetates with red dots on them could be placed. The red dots had a diameter of 8 mm, and there were 3 different acetates, with 1, 2, and 3 dots on them respectively.

A white screen was placed above the room's back wall to occlude the rest of the device from the participant's view. A 50-mm-diameter achromatic doublet lens of focal length 200 mm was placed behind this screen in line with the porthole and 255.5 mm from the position of the telescope mount. A blue screen, matched in color to the room, was placed 150 mm from this lens. As the red dots were placed on clear acetates, this blue screen acted as a background so the dots appeared as if they were on the back wall of the blue room.

Four telescopes were used, each comprising a 50-mm-diameter achromatic doublet lens attached to a 3-inch aluminum lens tube. There were two pairs of telescopes. In each pair, the invisible telescope had a focal length of 75 mm and the visible telescope had a focal length of 200 mm. To distinguish the telescopes in each pair, they were covered in yellow or green card. Telescope color was counterbalanced across participants.

The set-up of the cloaking device meant that when the visible telescope was placed on its mount in the blue room apparatus, it was possible to see the red dots against the blue background when looking through the telescope; whereas when the invisible telescope was in place, only the blue background was visible when looking through the telescope, the red dots were invisible (see Choi & Howell, 2014; Figure 1; Figure S.1; and Videos S.1 and S.2 in Supplemental Materials for details).

Computerized dot perspective task. The computer stimuli were based on those used in Santiesteban et al. (2014; Experiment

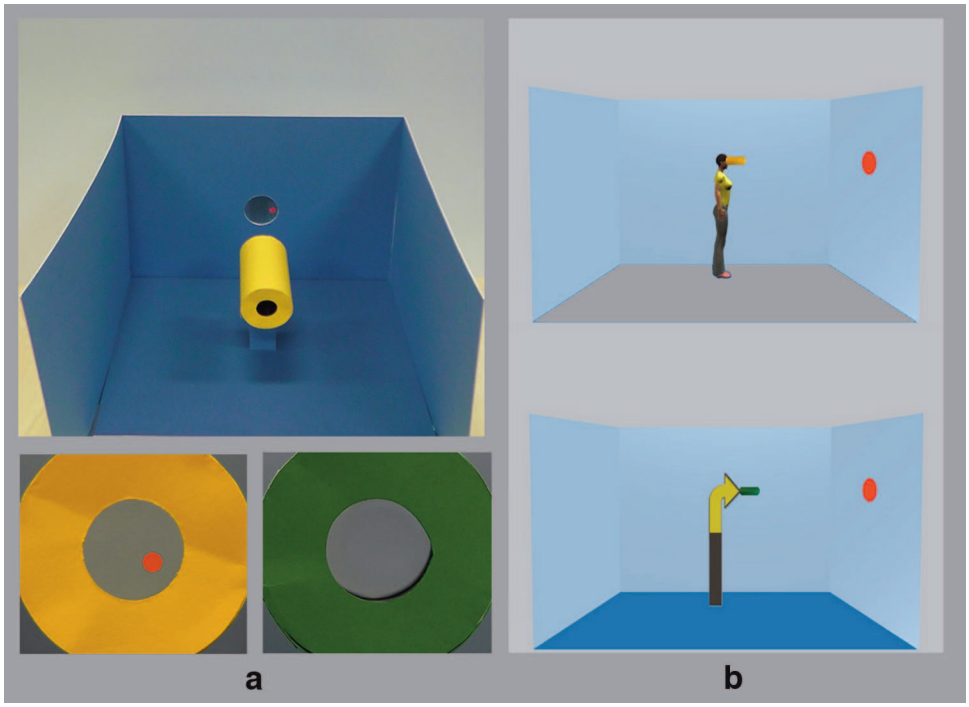

Figure 1. Examples of the cloaking device and computer stimuli in Experiment 1. Panel a (top) shows the blue room apparatus with one red dot present and that the red dot is seen through the visible telescope (panel a bottom left), but not the invisible telescope (panel a bottom right). Sample avatar and arrow stimuli with the telescopes for the computerized dot perspective task are depicted in panel b. See Supplemental Materials (Fig. S.1) and Choi and Howell (2014) for a full explanation of the invisibility effect. See the online article for the color version of this figure.

2), which were adapted from the original task images used by Samson et al. (2010; Experiment 3). A central stimulus was presented in the middle of a blue room facing either to the right or left. On some trials the stimulus was a human avatar and on others it was an arrow. The avatar and the arrow were matched in height, area, and color. There were two versions of each avatar and arrow: one 'male' and one 'female.' Participants viewed the central stimulus that matched their own gender. Our stimuli differed from Santiesteban et al. (2014; Experiment 2) in one respect: we inserted the green or yellow telescope into each image type (see Figure 1). On each trial, the green or yellow telescope appeared at the point of the arrow or at the eye of the avatar. Different configurations of red dots appeared on the front and back walls of the blue room. The possible number and configurations of dots were: 1 in front (F) or behind (B); 2F; 1F & 1B; 2B; 3F; 1F & 2B; 2F & 1B; 3B (Santiesteban et al., 2014). Participants completed the task on a laptop computer, and used the 'K' key (marked with a '1') to indicate a 'YES' response and the 'L' key (marked with a '2') to indicate a 'NO' response.

Procedure.

Telescope familiarization. Two telescopes of 200-mm focal length, one green and one yellow, were placed on a table in the testing room. The experimenter held up both telescopes and said, "Here are two telescopes, a green telescope and a yellow telescope. Take a look through them." At this stage, the two telescopes were of the same focal length so the difference in lens strength could not be detected. Participants could look around the testing room at anything they chose. After participants had examined each telescope, the experimenter asked them if they could see through each one, and then instructed them to carry both telescopes to the blue room apparatus which was situated in a separate cubicle within the testing room. Participants were asked to choose which telescope they would like to look through first. The experimenter placed the chosen telescope in the mount. The invisible telescope was covertly swapped for an identical telescope with a focal length of 75 mm (posttest debriefing revealed that no participant was aware of this switch). The experimenter then presented the 3 acetates with red dots to the participant for them to choose which order to view them in. The experimenter placed the first acetate on the back wall of the blue room and, while standing behind the participant, instructed them to look through the telescope. These steps were repeated for each of the 3 acetates for both the visible and the invisible telescope. Then, participants were asked to report what they thought the difference between the two telescopes was. This part of the procedure was video recorded. Following this, participants left the cubicle to complete the computerized task.

Dot perspective task. A fixation cross was shown (1250 ms) at the start of each trial, followed by the word 'YOU' (1250 ms) to indicate that the participant should judge how many dots they can see from their own perspective; then a number cue between 0 and 3 appeared (750 ms), followed by an image of the blue room. Participants were instructed to press '1' if the number cue matched the number of dots they could see in the image of the blue room, and '2' if the number cue did not match the number of dots they could see in the image. Participants were moved automatically onto the next trial once they made a response or after 2000 ms.

Apart from the inclusion of the telescope stimuli, the experimental design was the same as that used in Santiesteban et al. (2014; Experiment 2) and Samson et al. (2010; Experiment 3), with one further exception: the number of blocks was doubled to achieve the same number of trials per cell of the design as in these previous studies. Thus, participants completed 8 blocks of 52 trials each. Four trials in each block were filler trials in which no dots appeared. On half of the remaining trials, the avatar appeared and on half of these avatar trials the green telescope was present and on the other half the yellow telescope was present. The arrow was present on the remaining trials, half with the green telescope and half with the yellow. Half of the total nonfiller trials were 'inconsistent' and the other half 'consistent'; half required a 'YES' response and the other half a 'NO' response; and on half the central stimulus faced left and on the other half faced right. Trial types were therefore balanced across blocks. Trial order was pseudorandomized prior to testing to fulfill a rule that a similar trial type should not occur three consecutive times (Samson et al., 2010). Block order was randomized per participant. Participants first completed a practice block of 26 trials with accuracy feedback. No feedback was given on the experimental trials.

Manipulation check. Following the 8 experimental blocks, participants completed a further 12 trials in which they were presented with images of the blue room with the avatar stimulus. On half of these trials the yellow telescope was present and on the other half the green telescope was present; half of the trials were inconsistent and the other half were consistent. Different numbers and configurations of red dots were presented on each trial. Participants were asked to respond by pressing the keys 0/1/2/3 to indicate a response to the question: "*How many dots can the woman/man see through the green/yellow telescope*?"

Results

Analysis strategy. In keeping with previous studies, reaction time (RT) data were analyzed from 'YES' trials with correct responses only, using a 2 × 2 × 2 repeated measures ANOVA with within-subjects factors of Consistency (Consistent vs. Inconsistent), Stimulus (Avatar vs. Arrow), and Telescope Type (Visible vs. Invisible). The total number of errors was low ($M_{error\ rate}$ = 3%) and so accuracy data are reported in the Supplemental Materials; where effects are significant in the error data they are consistent with the RT data, providing no evidence for speed–accuracy trade-offs. Results were also analyzed within a Bayesian framework using JASP (https://jasp-stats.org; JASP Team, 2016), to examine the strength of the evidence in favor of the null and experimental hypotheses. Bayes Factors are particularly relevant to the current analyses as they provide a ratio of the likelihood of the observed data under the null versus alternative hypothesis, whereas p values examine the probability of the data given the null hypothesis and therefore cannot discriminate between evidence for the null and no evidence for either the null or alternative hypothesis (Dienes, 2016). Bayes Factors (BF01) are reported below, where values approaching zero indicate that the data provide more evidence in favor of the alternative hypothesis than the null hypothesis, a value of 1 indicates that the null and alternative hypotheses are equally likely given the data, and values above 1 indicate greater support for the null hypothesis. By convention values $<1/3$ and >3 are taken as evidence in favor of the alternative and null hypotheses, respectively, while values within these boundaries are judged to provide no evidence to favor either the null or alternative hypotheses.

Reaction time data. There was a main effect of Consistency, $F_{(1,42)} = 32.87$, $p < .001$, $\eta_p^2 = .439$, BF01 = 4.024×10^{-14}, whereby RTs (in ms) were significantly faster on consistent trials ($M = 514$, $SE = 15$, CI [483, 544]) than on inconsistent trials ($M = 549$, $SE = 19$, CI [511, 587]. There was also a main effect of Stimulus, $F_{(1,42)} = 4.39$, $p = .042$, $\eta_p^2 = .095$; RTs were significantly slower on trials on which the avatar was the central stimulus ($M = 535$, $SE = 17$, CI [500, 570]) rather than the arrow ($M = 528$, $SE = 16$, CI [495, 560]), but this was qualified by the fact that the Bayesian analysis found no support for either the null or alternative hypothesis (BF01 = 1.737). The Consistency x Stimulus interaction was also significant, $F_{(1,42)} = 6.89$, $p = .012$, $\eta_p^2 = .141$, but again the Bayesian analysis indicated no support for either the null or alternative hypothesis (BF01 = 1.182). It should also be noted that after controlling for the overall difference in RT between stimuli, this interaction was no longer significant ($F_{(1,41)} = 3.18$, $p = .08$, $\eta_p^2 = .072$). Crucially, a consistency effect was found for both the avatar stimulus ($F_{(1,42)} = 31.73$, $p < .001$, $\eta_p^2 = .430$, BF01 = 2.411×10^{-8}) and the arrow stimulus ($F_{(1,42)} = 21.84$, $p < .001$, $\eta_p^2 = .342$, BF01 = 8.056×10^{-5}). If the inanimate arrow stimulus can produce a consistency effect, then one cannot rely on the simple presence of a consistency effect as evidence for automatic mentalizing, as an arrow cannot 'see' the dots and does not have mental states.

Evidence of mentalizing would be obtained however, if the consistency effect varies as a function of Telescope Type for the avatar but not the arrow. The crucial statistics that would indicate evidence of automatic mentalizing are a significant 3-way interaction between Consistency × Stimulus × Telescope Type, or, less convincingly, a significant Consistency × Telescope Type interaction in the avatar condition only. No such evidence of automatic mentalizing was found however. The consistency effect did not vary as a function of Telescope Type and Stimulus (Consistency × Stimulus × Telescope Type: $F_{(1,42)} = 0.63$, $p = .43$, $\eta_p^2 = .015$, BF01 = 4.293), and in the avatar condition there was no effect of Telescope Type on the consistency effect (Consistency × Telescope Type: $F_{(1,42)} = 0.48$, $p = .49$, $\eta_p^2 = .011$, BF01 = 3.959) (means, standard errors, and 95% confidence intervals for these data are presented in Table S.1. in Supplemental Materials). As can be seen, the Bayes Factors provide support for the null over the alternative hypothesis in each case. Indeed, in the avatar condition, the consistency effect was numerically larger in the invisible telescope condition than in the visible telescope condition—a pattern opposite to that which would be predicted on the basis of automatic mentalizing (see Figure 2).

Confirmatory analysis. The logic of the telescope addition to the dot perspective task requires participants to be aware of the nature of each telescope. If participants should forget the fact that one telescope does not allow the red dots to be seen, or forget the mappings between telescope type and color, then it is possible that a Consistency × Stimulus × Telescope Type interaction would not be seen even if participants were automatically mentalizing. Accordingly, a very strict criterion was adopted such that only participants who correctly reported the difference between telescopes at the start of the experiment, and who responded correctly on 12 of 12 of the explicit questions at the end of the procedure ($n = 21$) were included. These participants were explicitly aware of the nature of the telescopes, and which telescope afforded seeing the red dots and which not, at the start and end of the experiment. Even among this highly selected set neither the Consistency × Stimulus × Telescope Type interaction ($F_{(1,20)} = 0.02$, $p = .88$, $\eta_p^2 = .001$, BF01 = 3.272), nor the Consistency × Telescope Type interaction within the avatar condition ($F_{(1,20)} = 0.02$, $p = .89$, $\eta_p^2 = .001$, BF01 = 3.273), was significant.

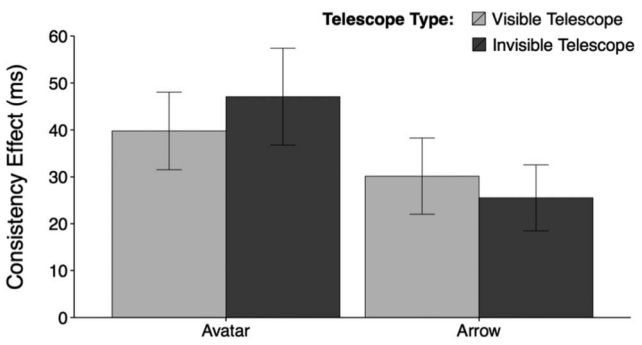

Figure 2. Mean consistency effect for each stimulus and telescope type in Experiment 1. Error bars show the standard error of the mean.

Discussion

The introduction of visible and invisible telescopes to the dot perspective taking task allowed a clear prediction to be made: if participants were automatically representing the avatar's mental state then a consistency effect should have been observed when the avatar was able to see through the visible telescope, but not when the avatar was faced with the invisible telescope, nor when the avatar was replaced with the arrow stimulus, regardless of which telescope accompanied it. Instead, a significant consistency effect was observed in all four conditions. Indeed, the consistency effect was numerically larger when the avatar looked through the invisible telescope than when it looked through the visible telescope, a pattern of data opposite to that predicted by the automatic mentalizing account. Although such a pattern of results is not consistent with the automatic mentalizing hypothesis, it is also inconsistent with the results obtained by Furlanetto et al. (2016). Experiments 2 and 3 investigate a potential explanation for this latter inconsistency.

Experiment 2

Experiment 1 found no evidence of automatic mentalizing in the dot perspective taking task using a visibility manipulation instantiated using a cloaking device to render the dots invisible. As outlined above, these results are in direct contrast to those obtained by Furlanetto et al. (2016) who used visible and invisible goggles to perform a conceptually similar experiment. We speculated that a possible reason for this discrepancy relates to the participants' task throughout the experiment. In Experiment 1 participants were required to verify whether the number cue matched the number of dots visible from their perspective only. In contrast, participants in the study of Furlanetto et al. were asked to respond on the basis of both their own and the avatar's perspective. This feature of the Furlanetto study makes it possible that effects of the avatar's perspective on own perspective trials were attributable to a task carry-over effect; that as a result of repeated demands to adopt the

avatar's perspective during the task, participants began to do so even on trials where it was not required. Experiment 2 tested for this possibility by implementing the Furlanetto procedure without avatar-perspective trials. Accordingly, participants were given experience of two pairs of goggles, one with transparent lenses through which they could see and the other with opaque lenses through which they could not see. Participants then completed the dot perspective task with the avatar stimulus both without goggles and with opaque and transparent goggles.

If the Furlanetto et al. (2016) effect is truly attributable to automatic mentalizing then one would expect the consistency effect to vary as a function of goggle type. Specifically, automatic mentalizing would be revealed by a consistency effect being observed when the avatar is wearing the transparent goggles and when wearing no goggles, but crucially not when wearing the opaque goggles. In contrast, if the modulation of the consistency effect by goggle type observed on own-perspective trials in the Furlanetto study was attributable to a task carry-over effect, then it should not be evident when participants respond on the basis of their own perspective only. Observation of a consistency effect in all three goggle conditions, including the crucial opaque condition, would indicate that the consistency effect in the dot perspective task is due to nonmentalistic domain-general processes, such as the directionality of the stimulus, and not automatic mentalizing.

Experiment 2 provided a further check on the generalizability of the results of Experiment 1. It could be argued that the cloaking device manipulation used in Experiment 1 is sufficiently novel, or sufficiently outside typical experience, that the automatic mentalizing system cannot represent the way in which it alters visual experience. Although the Confirmatory Analysis reported in Experiment 1 demonstrated that no sign of implicit mentalizing was observed in participants who we could be sure understood the visibility manipulation, the proportion of participants able to meet the strict understanding criterion used in this analysis was surprisingly low. The use in Experiment 2 of transparent and opaque goggles, stimuli with which participants are likely to have much greater experience, should alleviate the concern that the visibility manipulation is outside the realm in which the automatic mentalizing system can operate.

Method

Participants. Sixty-six healthy adults volunteered to take part in this experiment in return for a small monetary sum. Data from nine participants were excluded from the analysis; 4 because they did not follow instructions, and a further 5 for being outliers with respect to accuracy (error rate >25%). The remaining 57 participants (45 female) were aged between 18 and 56 ($M = 23.37$, $SD = 5.67$). The data-stopping rule and sample size were determined prior to data collection and were based on previous research. The target sample size was three times ($n = 54$) the size of the sample in Furlanetto et al. (2016; $n = 18$); see Simonsohn (2015) for discussion of the desirability of a sample at least 2.5 times that of the original study when attempting to replicate an effect.

The same participants completed Experiments 2 and 3. Experiment order was randomly assigned (Experiment 2 first: $n = 36$; Experiment 3 first: $n = 21$). As there were no effects of experiment order, and results for both experiments analyzed separately for each order were consistent with findings from the total sample, these samples were combined and data from the total sample are reported.

Stimuli and apparatus.

Goggles. Four pairs of goggles (two red and two orange) that matched the computerized stimuli from Furlanetto et al.'s (2016) study were used. The external lenses in all goggles were mirrored so that a person's eyes could not be seen through them. The internal lens in one red and one orange pair of goggles was covered with a blackout material so that they became opaque. The lenses in the other two pairs of goggles were unaltered and therefore remained transparent. The transparent and opaque goggles were indistinguishable when viewed externally.

Computerized dot perspective task. The computer stimuli were the exact same stimuli as those used in Furlanetto et al. (2016). The Furlanetto task was similar to that outlined in Experiment 1 except that: the room was gray and white with blue dots; the female avatar had a different physical appearance; the fixation cross and word cue were shown for 750 ms each with a 500-ms interstimulus interval; there was no arrow stimulus in this task; and the avatar appeared wearing either red, orange, or no goggles. Goggle type (transparent or opaque) was blocked, whereas whether the avatar was wearing goggles or no goggles was intermixed within a block, as per Furlanetto et al. Sample stimuli are depicted in Figure 3.

Procedure.

Belief induction. Participants were instructed on-screen before the transparent goggle condition, that "In this block the woman/man will sometimes wear orange/red goggles, so she/he will be able to see what is on the wall in front of her/him," or before the opaque goggle condition, that "so she/he will not be able to see what is on the wall in front of her/him." Following this, they were instructed "Now you will get first person experience of the

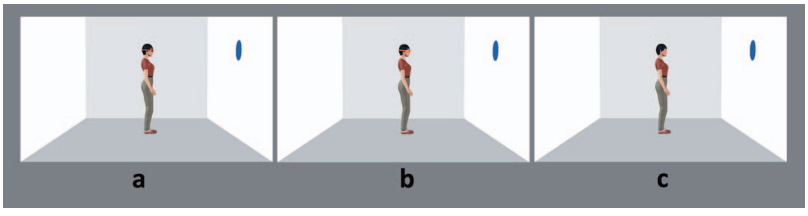

Figure 3. Examples of the computer stimuli in Experiments 2 and 3. Sample avatar stimuli from Furlanetto et al. (2016) with the red (panel a), orange (panel b), and no goggles (panel c) for the computerized dot perspective task in Experiments 2 and 3. See the online article for the color version of this figure.

visual experience of the woman/man." The experimenter then gave the participant the goggle type corresponding to the forthcoming condition (opaque or transparent) and asked the participant to look in the direction of the monitor for one minute. There were two separate belief inductions, one for each goggle type prior to the onset of both blocks for that condition (i.e., prior to the start of Block 1 and Block 3).

Dot perspective task. The presentation of the dot task was the same as that described in Experiment 1 except for the following changes. As Furlanetto et al. (2016) used 6 blocks of trials in total and had an additional factor that was not included in Experiment 2 (i.e., other-perspective trials on which participants had to respond based on the avatar's perspective), the current experiment used the self-perspective trials from their study (comprising 3 blocks of trials) and added an additional block to have an equal number for both the opaque and transparent goggles conditions (4 blocks in total).

There were 200 test trials in total, presented in four blocks of 50 trials (2 filler trials per block). Half of the trials in each block were consistent and the other half were inconsistent, half of the trials were matching (i.e., the number cue matched the number of dots participants could see in the image of the room) and the other half mismatching (i.e., the number cue did not match the number of dots participants could see in the image of the room). Within each block 33% of trials showed the avatar stimulus without any goggles and 66% of trials showed the avatar wearing either the red or orange goggles. In contrast to Experiment 1, goggle type (opaque or transparent) was never intermixed within blocks. Block order, opacity order, and goggle color were counterbalanced between participants.

As in Furlanetto et al. (2016), participants first completed 26 practice trials with feedback on trials on which the avatar stimulus had no goggles. No feedback was given on test trials. After the first belief induction phase participants completed the two blocks associated with that goggle condition, then received the second belief induction prior to commencing the final two blocks with the other goggle condition (e.g., two blocks with opaque goggles followed by two blocks with transparent goggles or vice versa). Between each of the four blocks participants were verbally reminded on-screen whether in the upcoming block the woman or man "will/will not be able to see what is on the wall in front of her/him."

Manipulation check. Participants were asked to choose a pair of goggles to wear while performing a visual search task. As in Furlanetto et al. (2016), all participants chose the transparent goggles.

Results

Analysis strategy. As in previous studies, RT data were analyzed from 'YES' trials with correct responses only, using a 2 × 3 repeated measures ANOVA with within-subjects factors of Consistency (Consistent vs. Inconsistent) and Goggle Type (No Goggles vs. Transparent Goggles vs. Opaque Goggles). The total number of errors was low ($M_{error\ rate}$ = 3.3%) and so accuracy data are reported in the Supplemental Materials; all significant effects in the error data are consistent with the RT data. Where sphericity assumptions were violated, Greenhouse-Geisser corrected values are reported.

Reaction time data. There was a main effect of Consistency, $F_{(1,56)} = 72.81$, $p < .001$, $\eta_p^2 = .565$, BF01 = 2.259×10^{-10}, whereby RTs (in ms) were significantly faster on consistent trials (M = 523, SE = 11, CI [501, 545]) than on inconsistent trials (M = 559, SE = 13, CI [534, 584]). There was no main effect of Goggle Type, $F_{(2,112)} = 1.90$, $p = .15$, $\eta_p^2 = .033$, BF01 = 3.751. The Consistency x Goggle Type interaction was significant, $F_{(2, 112)} = 4.58$, $p = .012$, $\eta_p^2 = .076$, BF01 = 1.378, although the Bayesian analysis provided no support for this effect. The Consistency x Goggle Type interaction was due to a significantly greater consistency effect in the Opaque (M = 49, SE = 8) than in the Transparent (M = 22, SE = 6) condition ($Mean_{diff}$ = 28, SE = 9, $p = .013$), whereas no other comparison was significant (Opaque vs. No Goggles: $Mean_{diff}$ = 14, SE = 9, $p = .424$; Transparent vs. No Goggles: $Mean_{diff}$ = −14, SE = 9, $p = .375$).

These data do not support an automatic mentalizing hypothesis, under which a consistency effect would be expected only in the conditions with the transparent goggles and no goggles. Indeed, the significantly greater consistency effect observed when the avatar wore Opaque versus Transparent goggles is directly opposite to what would be expected under the automatic mentalizing account. The lack of support for automatic mentalizing is evidenced by a significant consistency effect in all three conditions: when the avatar was wearing no goggles ($F_{(1,56)} = 34.36$, $p < .001$, $\eta_p^2 = .380$, BF01 = 2.435×10^{-5}), transparent goggles ($F_{(1,56)} = 13.12$, $p = .001$, $\eta_p^2 = .190$, BF01 = 0.025), and, crucially, opaque goggles ($F_{(1,56)} = 38.18$, $p < .001$, $\eta_p^2 = .405$, BF01 = 8.198×10^{-6}) (see Figure 4). Means, standard errors, and 95% confidence intervals for these data are presented in Table S.2. in Supplemental Materials.

Discussion

The results of Experiment 2 provide no support for the hypothesis that automatic mentalizing is responsible for the consistency effect in the dot perspective task. Instead, like the results of Experiment 1 and Cole et al. (2016), they are consistent with a submentalizing perspective in which domain-general processes such as attentional orienting underpin the consistency effect. Furthermore, alleviating any concerns that the cloaking device visibility manipulation in Experiment 1 was too novel or obtuse for the automatic mentalizing system to deal with, results were obtained

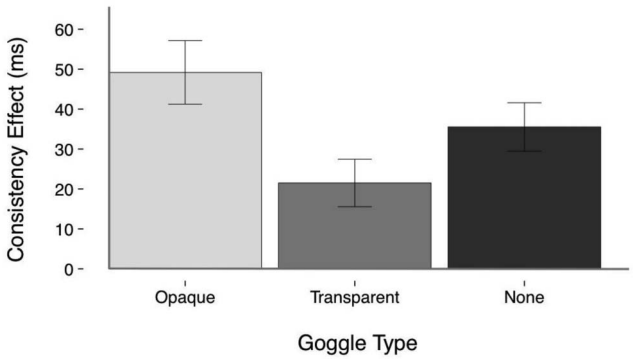

Figure 4. Mean consistency effect for each goggle type in Experiment 2. Error bars show the standard error of the mean.

with familiar materials in a familiar situation, and with explicit instructions as to the properties of the goggles.

Experiment 3

The results of Experiment 2 are consistent with the hypothesis that the positive evidence of automatic mentalizing reported by Furlanetto et al. (2016) is a task-specific product of a design in which participants are asked to adopt both their own perspective and that of the avatar. To further test this hypothesis, the participants from Experiment 2 also completed Experiment 3, which consisted of a straight replication of the Furlanetto et al. study including both self and avatar perspective trials. Comparison of the results of Experiment 2 and 3 will therefore enable the identification of a task carry-over effect should one exist. Evidence that the submentalizing process underpinning the consistency effect on self-perspective trials can be moderated by a carry-over effect of explicit, nonautomatic mentalizing on avatar-perspective trials would be demonstrated by the observation of a consistency effect in the crucial opaque goggles condition on self-perspective trials in Experiment 3.

Experiments 2 and 3 also investigated individual differences in the size of the consistency effect. A recent paper found that, on self-perspective trials, the consistency effect in the avatar condition was positively correlated with the perspective-taking and empathic concern subscales of the self-report 'Interpersonal Reactivity Index' questionnaire (IRI: Davis, 1983), whereas the consistency effect in an arrow, and a rectangular stimulus condition showed no such relationships (Nielsen, Slade, Levy, & Holmes, 2015). Nielsen et al. suggested that these results imply that consistency effects in the avatar condition reflect distinctly social processes that do not operate when consistency effects are observed with nonsocial stimuli. Participants in the current experiments (2 and 3) also completed the IRI to investigate whether the consistency effect in the avatar condition varies according to empathy and perspective-taking.

Method

The method for Experiment 3 was the same as that for Experiment 2 with the following exceptions described below and was an exact replication of that of Furlanetto et al. (2016), using the same stimuli, design, and procedure.

Dot perspective task. There were 300 test trials presented in six blocks of 50 trials (including 2 filler trials per block). There were 3 blocks per goggle condition. There was an additional factor of Perspective with 2 levels, Self and Other. Half of all trials were Self and the other half were Other trials. The word cue 'YOU' indicated that the participant should judge how many dots they can see from their own perspective (as in Experiments 1 and 2), the word cue 'SHE/HE' indicated that the participant should judge how many dots the avatar can see from the avatar's perspective. Self and Other trials were intermixed within blocks.

Interpersonal reactivity index. On completion of the study, participants were asked to complete the perspective-taking (PT) and empathic concern (EC) subscales of the IRI via an online link to the questionnaire. Each subscale had 7-items scored on a 5-point Likert scale (0 = 'does not describe me well'; 4 = 'describes me very well'), and measured the tendency to adopt others' point of view (PT: $\alpha = .72$), and have concern or compassionate feelings for others (EC: $\alpha = .78$; Davis, 1983). Sample items included "when I am upset at someone, I usually try to 'put myself in his shoes' for a while" (PT), and "I would describe myself as a pretty soft-hearted person" (EC).

Results

Analysis strategy. As in previous studies, RT data were analyzed from 'YES' trials with correct responses only, using a 2 × 2 × 3 repeated measures ANOVA with within-subjects factors of Consistency (Consistent vs. Inconsistent), Perspective (Self vs. Other), and Goggle Type (No Goggles vs. Transparent Goggles vs. Opaque Goggles). The total number of errors was low ($M_{error\ rate}$ = 6.4%) and so accuracy data are reported in the Supplemental Materials; where effects are significant in the error data they are consistent with the RT data. Where sphericity assumptions were violated, Greenhouse-Geisser corrected values are reported.

The relationship between the consistency effects and the perspective-taking and empathic concern subscales of the IRI were examined as in the study by Nielsen et al. (2015), using one-tailed Pearson's correlations, and using Bayesian correlations. Analyses were conducted for both overall consistency effects and by goggle type for the self-perspective trials from Experiment 2 and Experiment 3 separately, and the other-perspective trials from Experiment 3.

Reaction time data. There was a main effect of Consistency, $F_{(1,56)} = 75.01$, $p < .001$, $\eta_p^2 = .573$, BF01 = 1.029×10^{-11}, whereby RTs (in ms) were significantly faster on consistent trials ($M = 602$, $SE = 15$, CI [573, 632]) than on inconsistent trials ($M = 660$, $SE = 19$, CI [622, 697]). There was also a main effect of Perspective, $F_{(1,56)} = 39.69$, $p < .001$, $\eta_p^2 = .415$, BF01 = 2.575×10^{-9}, with faster responding on Self trials ($M = 605$, $SE = 16$, CI [573, 637]) than on Other trials ($M = 657$, $SE = 18$, CI [621, 692]). The Consistency × Perspective interaction was significant, $F_{(1,56)} = 11.22$, $p = .001$, $\eta_p^2 = .167$, BF01 = 0.064, with a larger consistency effect in the Other condition ($M = 80$, $SE = 10$, CI [61, 99]) than the Self condition ($M = 35$, $SE = 9$, CI [16, 53]). There was a significant Consistency × Perspective × Goggle Type interaction, $F_{(1.6,\ 92.2)} = 4.86$, $p = .015$, $\eta_p^2 = .080$, BF01 = 0.631, which was not supported by the Bayesian analysis.

The Consistency × Perspective × Goggle Type interaction was driven by a Consistency × Goggle Type interaction that was significant in the Other condition, $F_{(2,\ 112)} = 6.042$, $p = .003$, $\eta_p^2 = .097$, BF01 = 0.285, driven in turn by the fact that the consistency effect in the Opaque Goggles Other condition was significantly smaller than in the Transparent Goggles Other condition, $F_{(1,\ 56)} = 5.74$, $p = .020$, $\eta_p^2 = .093$, BF01 = 1.140, and than in the No Goggles Other condition, $F_{(1,\ 56)} = 9.18$, $p = .004$, $\eta_p^2 = .141$, BF01 = 0.376, although neither of these simple contrasts were supported by the Bayesian analysis (see Figure 5). The difference in the size of the consistency effect between the Transparent Goggles Other and No Goggles Other conditions was not significant, $F_{(1,\ 56)} = 0.555$, $p = .459$, $\eta_p^2 = .010$, BF01 = 4.241.

The consistency effect observed on such Other perspective trials, that is, when judging the avatar's perspective, is an example of egocentric intrusion (Samson et al., 2010). On these trials the participant is explicitly instructed to adopt the avatar's perspective and they are slower to do so when the avatar's perspective is

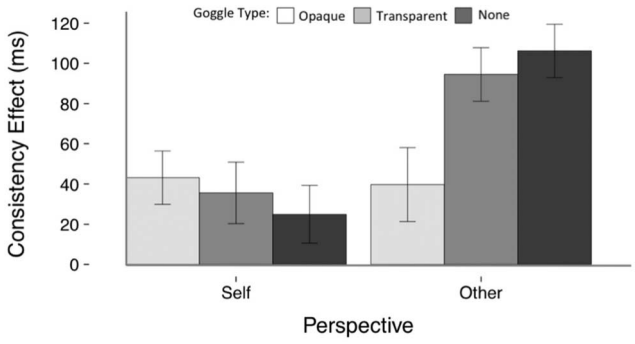

Figure 5. Mean consistency effect for each perspective and goggle type in Experiment 3. Error bars show the standard error of the mean.

inconsistent with their own than when it is consistent. Although this effect is interesting, it does not bear upon whether the avatar's perspective is automatically represented on self-perspective trials. The reduction in the egocentric intrusion effect with opaque goggles is encouraging however, as it suggests that, when explicitly instructed to adopt the avatar's perspective, participants were representing that the avatar could not see any dots when wearing the opaque goggles. Therefore what were previously 'consistent' trials were now in fact inconsistent, as when wearing opaque goggles the avatar never saw any dots on the wall it was facing whereas the participant did on all 'consistent' trials that were not filler trials (note that on filler trials no dots appeared; 12/300 trials were fillers and these were not analyzed). Therefore for all of the 'consistent' trials analyzed in the opaque goggle condition, the avatar's and participant's perspectives were conflicting, thus slowing responding.

In contrast, on Self-Perspective trials the Consistency × Goggle Type interaction was not significant, $F_{(2,\ 112)} = 0.48$, $p = .619$, $\eta_p^2 = .009$, BF01 = 12.643. The consistency effect did not vary as a function of goggle type, and there was a significant consistency effect in the Opaque Goggles Self condition, $F_{(1,\ 56)} = 10.58$, $p = .002$, $\eta_p^2 = .159$, BF01 = 0.064, in the Transparent Goggles Self condition, $F_{(1,\ 56)} = 5.434$, $p = .023$, $\eta_p^2 = .088$, BF01 = 0.505, and a marginally significant effect in the No Goggles Self condition, $F_{(1,\ 56)} = 3.00$, $p = .089$, $\eta_p^2 = .051$, BF01 = 1.398 (see Figure 5). Note that of the consistency effects in the individual conditions, only that in the Opaque Self condition was supported by the Bayesian analysis, and the Bayesian analysis provided strong support for the lack of any effect of Goggle Type on consistency. Means, standard errors, and 95% confidence intervals for these data are presented in Table S.3. in Supplemental Materials.

The self-perspective trials from Experiment 3 and Experiment 2 (which included only self-perspective trials) were compared to examine whether the Consistency × Goggle Type interaction varied as a function of Experiment. However, the Consistency × Goggle Type × Experiment interaction was not significant, $F_{(2,\ 112)} = 0.762$, $p = .469$, $\eta_p^2 = .013$, BF01 = 11.923, providing no evidence that task carry-over effects influence the consistency effect.

Interpersonal reactivity index. Forty-five participants responded to the questionnaire. In the full data set, there were no significant correlations (all ps $> .05$). In a reduced data set ($n = 35$), from which outliers were removed using the 1.5 interquartile range rule (Tukey, 1977), the only significant correlation observed prior to correcting for multiple testing was a positive relationship between empathic concern and the consistency effect on other-perspective trials in the no goggles condition, $r = .35$, $p = .02$. The interpretation of this correlation is unclear, as in the conceptually similar transparent goggles condition (in which the avatar can also 'see') it was not observed, $r = -.04$, $p = .41$. After correcting for multiple comparisons it no longer reached significance. Bayesian analyses showed no support for any correlations in both the full and reduced data set. It is clear, therefore, that these results do not support those observed by Nielsen et al. (2015).

Discussion

Experiment 3 represented a replication of Furlanetto et al. (2016) with greater power to detect any effects, if present. Despite this, it was not possible to replicate the original results; the magnitude of the consistency effect on self-perspective trials was not modulated as a function of whether the avatar was wearing transparent or opaque goggles. In contrast, strong evidence was obtained from the Bayesian analysis of these data that the consistency effect was *not* modulated by which goggles the avatar was wearing. That the avatar's visual perspective was manipulated without any effect on participants' responding on self-perspective trials indicates that automatic representation of the avatar's mental state does not generate the consistency effect.

The cross-experimental comparison, demonstrating that the consistency effect by goggle type interaction on self-perspective trials is similar in both the self-perspective-only experiment (2) and in the self condition from the mixed-perspective experiment (3), suggests that in Experiment 3, the consistency effect on self-perspective trials was not affected by task carry-over effects from other-perspective trials. Given our larger sample size, our results are therefore more consistent with those from Furlanetto et al. (2016) reflecting a false positive, rather than being attributable to carry-over effects.

In contrast to the findings of Nielsen et al., these data showed no relationship between empathic concern or perspective-taking and the consistency effect. The data therefore do not support the claim that consistency effects for avatar stimuli involve specific mentalistic, or general social, processes.

General Discussion

The novel invisibility manipulation used in Experiment 1 allowed us to develop an experimental paradigm in which, should evidence consistent with automatic mentalizing have been found, one could reasonably claim that a submentalizing process could not have been responsible for the observed results. In contrast, we found no evidence that participants were automatically representing what the avatar can see in the dot perspective task. Whether the avatar was looking through a telescope through which they either could, or could not, see the red dots made no difference to the size of the consistency effect, a finding which runs counter to any explanation of the consistency effect being due to the representation of what the avatar can see. Similarly, replicating Santiesteban et al. (2014), a consistency effect was also observed for the arrow stimulus. Furthermore, our reexamination of the design and pro-

cedure used by Furlanetto et al. (2016) found no support for the claim that ascription of mental states underpins the consistency effect, nor for the possibility that this effect could be modulated by a task carry-over effect of explicit mentalizing (Experiments 2 and 3). Together these findings suggest that domain-general nonmentalistic processes, such as automatic directional cueing, underpin the consistency effect previously found using the dot perspective task.

The current Experiments 2 and 3 also showed no relationship between the size of the consistency effect and individual differences in empathic concern or perspective-taking, and therefore do not support the suggestion by Nielsen et al. (2015) that consistency effects in the avatar condition reflect distinctly social processes. As further support of this claim, Nielsen et al. also pointed to a significantly larger consistency effect on self-perspective trials in the avatar (i.e., social) condition compared with two nonsocial conditions (an arrow and a rectangle). However, the avatar stimulus was significantly larger than the arrow and rectangle stimuli, which were comparable in size, and therefore it is possible that the larger consistency effect in the avatar condition was a result of the size of the central stimulus rather than its social aspects. This confound, and the lack of replication in the current experiments, suggests that processes underlying the consistency effect are not social in nature.

In the automatic (or 'implicit') mentalizing literature, a distinction is often made between 'Level 1' and 'Level 2' perspective taking, where Level 1 refers to the ability to "infer what object another person does and does not see" (Flavell, Abrahams Everett, Croft, & Flavell, 1981, p. 99), and Level 2 refers to knowing "that an object simultaneously visible to both the self and the other person may nonetheless give rise to different visual impressions or experiences in the two if their viewing circumstances differ" (Flavell et al., 1981, p.99). Level 1 perspective taking thus concerns the visibility of an object, while Level 2 perspective taking concerns its appearance. It has been claimed that the automatic and efficient route to belief or belief-like state representation is limited to Level 1 perspective taking only (Apperly & Butterfill, 2009; Butterfill & Apperly, 2013; Qureshi et al., 2010; Surtees, Butterfill, & Apperly, 2012). The dot task is a measure of Level 1 perspective taking as, under the mentalizing account, the consistency effect depends on inferring that the avatar does see the dots on the wall in front of them but does not see the dots on the wall behind them (Apperly & Butterfill, 2009; Furlanetto et al., 2016; Qureshi et al., 2010; Samson et al., 2010; Surtees, Butterfill, & Apperly, 2012).

Crucially, the introduction of a visibility manipulation, as in the current studies and in the studies by Cole et al. (2016) and Furlanetto et al. (2016), does not alter the level of perspective taking of the dot task; rather, it manipulates Level 1 perspective taking: whether another person can see an object seen by oneself. The invisible telescope does not change the appearance of the dots in a way that would qualify for Level 2 perspective taking (e.g., by making them change color while remaining jointly visible to both avatar and participant). The invisible telescope changes the *visibility* of the dots, not their *appearance,* therefore allowing a manipulation of Level 1 perspective taking.

The current experiments add to an emerging literature that reexamines claims of automatic mentalizing as a domain-specific process of mental state representation (Phillips et al., 2015; Santiesteban et al., 2014; Schneider, Lam, Bayliss, & Dux, 2012). A recent reexamination (Phillips et al., 2015) of a different task, first used to support claims of automatic mentalizing in adults and 7-month-old infants (Kovács et al., 2010), demonstrated that the observed effects result from an attention-check rather than automatic mentalizing. The current Experiment 1 goes beyond the analysis of existing effects however, by providing a manipulation by which automatic mentalizing could be detected, if present. Even if it were possible that automatic mentalizing might occur but not interfere with the dot task, the current experiments invalidate the mentalizing interpretation of the consistency effect, showing it is not caused by interference from spontaneous computation of the avatar's conflicting visual perspective.

The finding that mental states are not necessarily represented in tasks putatively assumed to measure automatic mentalizing has profound implications. Evidence of automatic mentalizing has been used in support of claims including its evolutionary significance as a uniquely human adaptation (Kovács et al., 2010), specific deficits in those with Autism Spectrum Disorder (Senju, Southgate, White, & Frith, 2009), and the presence of a dual-process system for mentalizing (Apperly & Butterfill, 2009; Apperly, 2011). These data suggest that mentalizing may not be as pervasive as previously assumed (Apperly, 2011).

Our findings also contribute to the intriguing possibility that what has been termed 'automatic mentalizing' might in fact be entirely accounted for by domain-general processes and, although someone may act *as if* they understand another person's mental state, no mental states are actually represented (Heyes, 2014a). This opens up new avenues for research to investigate how cultural learning may underpin the development of a full-blown *explicit* mentalizing ability, what ontogenetic experiences enhance or impair this ability, and what factors, such as motivation or intelligence, influence individual differences in the degree of mentalizing skill and the degree to which this skill is applied in everyday life.

References

Apperly, I. A. (2011). *Mindreaders: The cognitive basis of "Theory of Mind."* Hove, UK: Psychology Press.

Apperly, I. A., & Butterfill, S. A. (2009). Do humans have two systems to track beliefs and belief-like states? *Psychological Review, 116,* 953–970. http://dx.doi.org/10.1037/a0016923

Butterfill, S. A., & Apperly, I. A. (2013). How to construct a minimal theory of mind. *Mind & Language, 28,* 606–637. http://dx.doi.org/10.1111/mila.12036

Catmur, C., Santiesteban, I., Conway, J. R., Heyes, C., & Bird, G. (2016). Avatars and arrows in the brain. *NeuroImage, 132,* 8–10. http://dx.doi.org/10.1016/j.neuroimage.2016.02.021

Choi, J. S., & Howell, J. C. (2014). Paraxial ray optics cloaking. *Optics Express, 22,* 29465–29478. http://dx.doi.org/10.1364/OE.22.029465

Cole, G. G., Atkinson, M., Le, A. T. D., & Smith, D. T. (2016). Do humans spontaneously take the perspective of others? *Acta Psychologica, 164,* 165–168. http://dx.doi.org/10.1016/j.actpsy.2016.01.007

Davis, M. H. (1983). Measuring individual differences in empathy: Evidence for a multidimensional approach. *Journal of Personality and Social Psychology, 44,* 113–126. http://dx.doi.org/10.1037/0022-3514.44.1.113

Dienes, Z. (2016). How Bayes factors change scientific practice. *Journal of Mathematical Psychology, 72,* 78–89. http://dx.doi.org/10.1016/j.jmp.2015.10.003

Flavell, J. H., Everett, B. A., Croft, K., & Flavell, E. R. (1981). Young children's knowledge about visual perception: Further evidence for the Level 1–Level 2 distinction. *Developmental Psychology, 17,* 99–103. http://dx.doi.org/10.1037/0012-1649.17.1.99

Furlanetto, T., Becchio, C., Samson, D., & Apperly, I. (2016). Altercentric interference in level 1 visual perspective taking reflects the ascription of mental states, not submentalizing. *Journal of Experimental Psychology: Human Perception and Performance, 42,* 158–163. http://dx.doi.org/10.1037/xhp0000138

Heyes, C. M. (1998). Theory of mind in nonhuman primates. *Behavioral and Brain Sciences, 21,* 101–114. http://dx.doi.org/10.1017/S0140525X98000703

Heyes, C. (2014a). Submentalizing: I am not really reading your mind. *Perspectives on Psychological Science, 9,* 131–143. http://dx.doi.org/10.1177/1745691613518076

Heyes, C. (2014b). False belief in infancy: A fresh look. *Developmental Science, 17,* 647–659. http://dx.doi.org/10.1111/desc.12148

Heyes, C. (2015). Animal mindreading: What's the problem? *Psychonomic Bulletin & Review, 22,* 313–327. http://dx.doi.org/10.3758/s13423-014-0704-4

JASP Team (2016). JASP (Version 0.7.5.5)[Computer software].

Kovács, Á. M., Téglás, E., & Endress, A. D. (2010). The social sense: Susceptibility to others' beliefs in human infants and adults. *Science, 330,* 1830–1834. http://dx.doi.org/10.1126/science.1190792

Nielsen, M. K., Slade, L., Levy, J. P., & Holmes, A. (2015). Inclined to see it your way: Do altercentric intrusion effects in visual perspective taking reflect an intrinsically social process? *Quarterly Journal of Experimental Psychology, 68,* 1931–1951. http://dx.doi.org/10.1080/17470218.2015.1023206

Penn, D. C., & Povinelli, D. J. (2007). On the lack of evidence that non-human animals possess anything remotely resembling a 'theory of mind'. *Philosophical Transactions of the Royal Society of London Series B, Biological Sciences, 362,* 731–744. http://dx.doi.org/10.1098/rstb.2006.2023

Phillips, J., Ong, D. C., Surtees, A. D., Xin, Y., Williams, S., Saxe, R., & Frank, M. C. (2015). A second look at automatic theory of mind: Reconsidering Kovács, Téglás, and Endress (2010). *Psychological Science, 26,* 1353–1367. http://dx.doi.org/10.1177/0956797614558717

Qureshi, A. W., Apperly, I. A., & Samson, D. (2010). Executive function is necessary for perspective selection, not Level-1 visual perspective calculation: Evidence from a dual-task study of adults. *Cognition, 117,* 230–236. http://dx.doi.org/10.1016/j.cognition.2010.08.003

Samson, D., Apperly, I. A., Braithwaite, J. J., Andrews, B. J., & Bodley Scott, S. E. (2010). Seeing it their way: Evidence for rapid and involuntary computation of what other people see. *Journal of Experimental Psychology: Human Perception and Performance, 36,* 1255–1266. http://dx.doi.org/10.1037/a0018729

Santiesteban, I., Catmur, C., Hopkins, S. C., Bird, G., & Heyes, C. (2014). Avatars and arrows: Implicit mentalizing or domain-general processing? *Journal of Experimental Psychology: Human Perception and Performance, 40,* 929–937. http://dx.doi.org/10.1037/a0035175

Schneider, D., Lam, R., Bayliss, A. P., & Dux, P. E. (2012). Cognitive load disrupts implicit theory-of-mind processing. *Psychological Science, 23,* 842–847. http://dx.doi.org/10.1177/0956797612439070

Schurz, M., Kronbichler, M., Weissengruber, S., Surtees, A., Samson, D., & Perner, J. (2015). Clarifying the role of theory of mind areas during visual perspective taking: Issues of spontaneity and domain-specificity. *NeuroImage, 117,* 386–396. http://dx.doi.org/10.1016/j.neuroimage.2015.04.031

Senju, A., Southgate, V., Snape, C., Leonard, M., & Csibra, G. (2011). Do 18-month-olds really attribute mental states to others? A critical test. *Psychological Science, 22,* 878–880. http://dx.doi.org/10.1177/0956797611411584

Senju, A., Southgate, V., White, S., & Frith, U. (2009). Mindblind eyes: An absence of spontaneous theory of mind in Asperger syndrome. *Science, 325,* 883–885. http://dx.doi.org/10.1126/science.1176170

Simonsohn, U. (2015). Small telescopes: Detectability and the evaluation of replication results. *Psychological Science, 26,* 559–569. http://dx.doi.org/10.1177/0956797614567341

Surtees, A. D., Butterfill, S. A., & Apperly, I. A. (2012). Direct and indirect measures of Level-2 perspective-taking in children and adults. *British Journal of Developmental Psychology, 30,* 75–86. http://dx.doi.org/10.1111/j.2044-835X.2011.02063.x

Tukey, J. W. (1977). *Exploratory data analysis*. Reading, MA: Addison-Wesley.

Received May 25, 2016
Revision received August 15, 2016
Accepted August 29, 2016 ■

6.8 Supplemental Materials

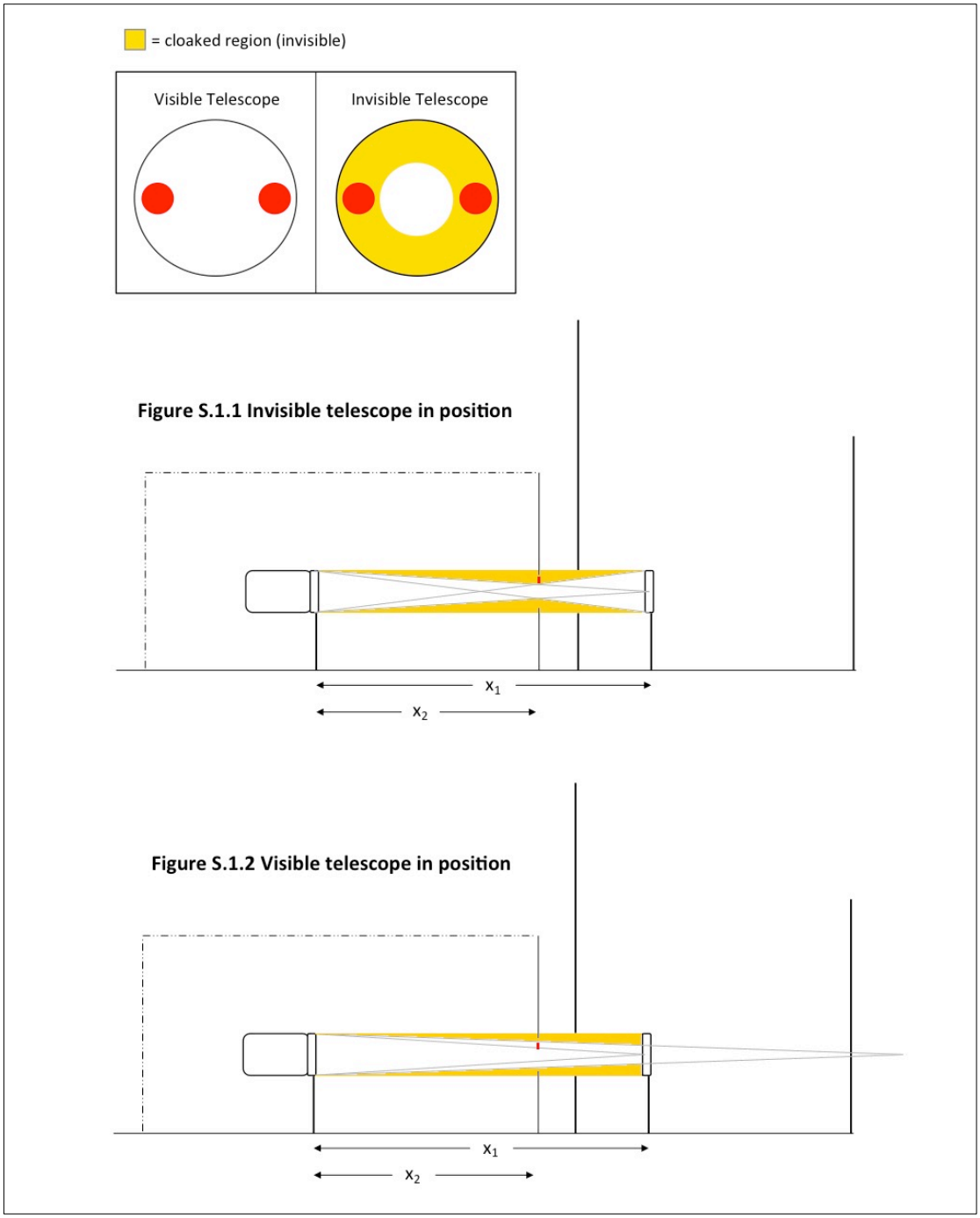

Figure 6.6 Schematic diagram of the cloaking device in Experiment 1. A diagram of the cloaking device. The dashed line represents the outline of the blue room. Participants looked horizontally through the system from the left hand side of the diagram. Distances x_1 = 255.5mm and x_2 = 150mm. The telescope was placed on a

mount. A high white screen was situated behind the back of the blue room so that the remaining apparatus was occluded from the participant's view. A 45mm diameter circular hole was cut into this white screen and the back wall of the blue room. Transparent acetates with opaque red dots were placed on the back wall of the blue room so that they appeared within this circle. A blue screen was situated at the end of the system to act as the background when looking through the system. Figure 6.6 S.1.1 shows the region cloaked by the invisible telescope (75mm focal length). The red dot falls within the cloaked region when viewed through the invisible telescope and therefore cannot be seen. Figure 6.6 S.1.2 shows the region cloaked by the visible telescope (200mm focal length). The red dot does not fall within the cloaked region when viewed through the visible telescope and therefore can be seen (see also the videos at https://osf.io/jas4n/ and Choi & Howell, 2014 for further details).

6.8.1 Reaction Time Data

	Consistent			Inconsistent		
	M	SE	95% *CI*	M	SE	95% *CI*
Telescope Type	Avatar					
Visible	514	15	[483, 545]	554	20	[514, 594]
Invisible	512	16	[481, 544]	560	22	[516, 604]
	Arrow					
Visible	512	14	[484, 541]	543	19	[504, 581]
Invisible	515	17	[482, 549]	541	17	[506, 576]

Table 6.1 Experiment 1 Means (M), Standard Errors (SE) and 95% Confidence Intervals (CI) for Reaction Time Data, in milliseconds, for each Trial Type.

	Consistent			Inconsistent		
	M	SE	95% *CI*	M	SE	95% *CI*
Goggle Type	Self					
Opaque	512	10	[491, 533]	561	15	[532, 591]
Transparent	538	13	[513, 564]	560	14	[532, 587]
None	520	12	[496, 544]	555	13	[530, 581]

Table 6.2 Experiment 2 Means (M), Standard Errors (SE) and 95% Confidence Intervals (CI) for Reaction Time Data, in milliseconds, for each Trial Type.

	Consistent			Inconsistent		
	M	SE	95% *CI*	M	SE	95% *CI*
Goggle Type	Self					
Opaque	589	19	[551, 626]	632	22	[588, 675]
Transparent	584	14	[556, 612]	620	22	[576, 663]
None	591	17	[558, 624]	616	20	[577, 656]
	Other					
Opaque	650	26	[598, 701]	690	30	[630, 749]
Transparent	589	15	[559, 620]	684	19	[647, 721]
None	611	15	[580, 641]	717	20	[677, 757]

Table 6.3 Experiment 3 Means (M), Standard Errors (SE) and 95% Confidence Intervals (CI) for Reaction Time Data in milliseconds, for each Trial Type.

6.8.2 Accuracy Data

	F	p	η_p^2	BF01
Consistency	21.61	< .001	.278	2.295×10^{-8}
Perspective	24.26	< .001	.302	3.675×10^{-6}
Goggle Type	7.10	.001	.112	0.052
Consistency x Perspective	0.17	.682	.003	8.629
Consistency x Goggle Type	13.40	< .001	.193	0.037
Perspective x Goggle Type	5.11	.007	.084	0.530
Consistency x Perspective x Goggle Type	14.14	< .001	.202	0.007

Table 6.4 Experiment 1 Accuracy Results from Repeated Measures ANOVA with Factors Consistency, Stimulus, and Telescope Type.
Consistency x Telescope Type in the Avatar condition:
$F_{(1,42)} = 1.40$ $p = .244$, $\eta_p^2 = .032$, BF01 = 2.308;
and in the Arrow condition:
$F_{(1,42)} = 2.54$, $p = .119$, $\eta_p^2 = .057$, BF01 = 4.152.

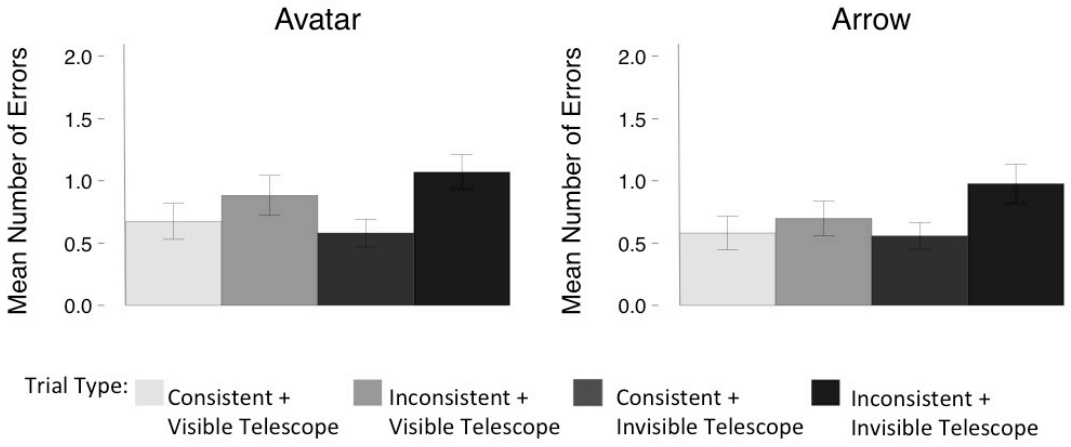

Figure 6.7 Experiment 1 Mean Number of Errors for Each Consistency, Stimulus and Telescope Type.
Error bars show the Standard Error of the Mean.

	F	*p*	η_Q^2	*BF01*
Consistency	10.58	.002	.159	0.020
Goggle Type	0.633	.516	.011	19.485
Consistency x Goggle Type	3.323	.040	.056	0.864

Table 6.5 Experiment 2 Accuracy Results from Repeated Measures ANOVA with Factors Consistency, and Goggle Type.
Consistency x Goggle Type for Opaque vs Transparent Goggles:
$F_{(1,56)} = 2.603, p = .112, \eta_Q^2 = .044$, BF01 = 1.382;
Opaque vs No Goggles:
$F_{(1,56)} = 6.871, p = .011, \eta_Q^2 = .109$, BF01 = 2.405;
and Transparent vs No Goggles:
$F_{(1,56)} = 0.665, p = .418, \eta_Q^2 = .012$, BF01 = 1.973.

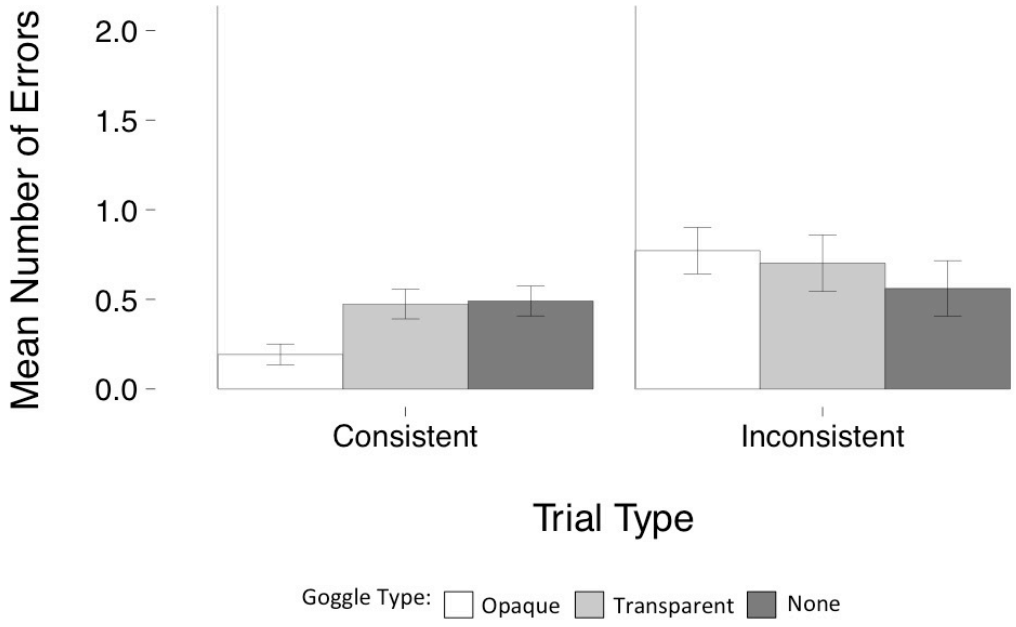

Figure 6.8 Experiment 2 Mean Number of Errors for Each Consistency and Goggle Type.
Error bars show the Standard Error of the Mean.

	F	p	η_p^2	BF01
Consistency	21.61	< .001	.278	2.295×10^{-8}
Perspective	24.26	< .001	.302	3.675×10^{-6}
Goggle Type	7.10	.001	.112	0.052
Consistency x Perspective	0.17	.682	.003	8.629
Consistency x Goggle Type	13.40	< .001	.193	0.037
Perspective x Goggle Type	5.11	.007	.084	0.530
Consistency x Perspective x Goggle Type	14.14	< .001	.202	0.007

Table 6.6 Experiment 3 Accuracy Results from Repeated Measures ANOVA with Factors Consistency, Perspective, and Goggle Type.
Consistency x Goggle Type in the Self condition:
$F_{(2,112)} = 0.828, p = .440, \eta_p^2 = .015, BF01 = 10.545;$
and in the Other condition: Consistency X Goggle Type:
$F_{(2,112)} = 19.84, p < .001, \eta_p^2 = .262, BF01 = 4.621 \times 10^{-4}.$

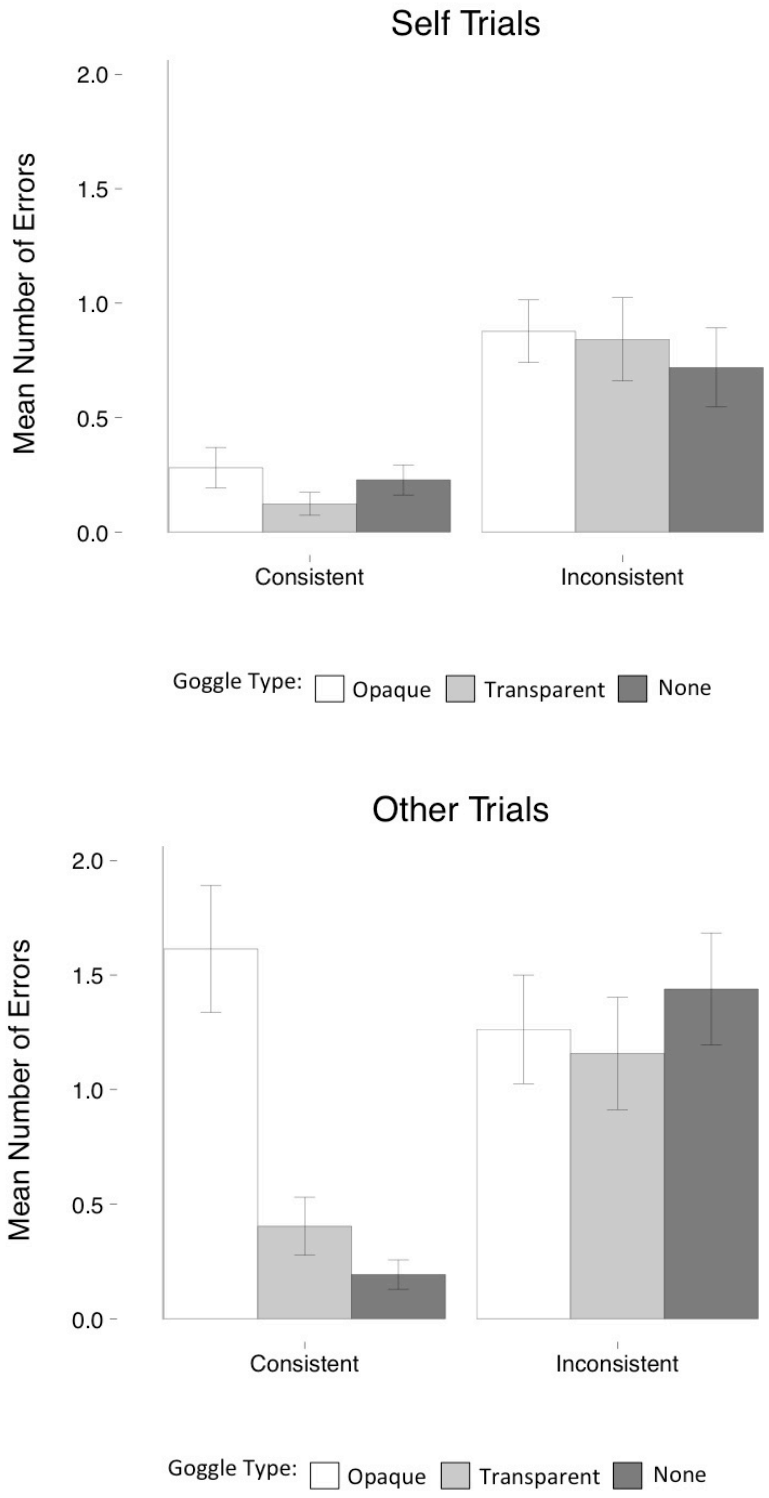

Figure 6.9 Experiment 3 Mean Number of Errors for Each Consistency, Perspective, and Goggle Type.
Error bars show the Standard Error of the Mean.

7. General Discussion

7.1 Thesis Summary

This thesis aimed to develop a new theoretical framework for conceptualising individual differences in theory of mind. Chapter 2 addressed why variance has proved difficult to measure in existing tasks beyond capturing the acquisition of specific mental state concepts, and correlations with general cognitive abilities and social experience. It also proposed that variance between minds in a population is an important factor in explaining variance in what mental states such minds would hold. Chapter 2 introduced the Mind-space framework for explaining and studying individual differences in mind representation and mental state inference, and further considered its relationship to existing dimensional theories (Section 2.9), and its implications for understanding typical and atypical development (Section 2.11). Notably, it demonstrated how Mind-space could generate testable predictions, some of which were investigated in Chapters 3-5.

Chapter 3 presented four experiments that found empirical support for the Mind-space theory. Experiment 1 showed that the more accurately an individual represented the population covariance between six personality dimensions, the more accurate their mental state inferences on an existing theory of mind task. Experiment 2 showed that those with a more accurate Mind-space covariance structure were more accurate at locating individual minds in Mind-space, based only on a thin-slice sample of their behaviour. Experiment 2 also found that the more similar the participant was to the target mind, the more accurate they were at locating that mind in Mind-space. Experiment 3 demonstrated that the probability of inferring a particular mental state depended on the dimensional location of the target mind in Mind-space. This finding was further supported by Experiment 4, which also showed that mental state probability

depended on the dimensional location of the target mind. In this case, the location was inferred from the dimension's covariance with six other dimensions. Chapter 3 thus provided evidence of how representing variance in minds relates to variance in mental state inferences.

Chapter 4 examined whether the mapping of minds onto trait dimensions in Mind-space changes as a result of experiencing populations with different trait distributions, and attempted to investigate this using adaptation experiment designs from the perception literature. No significant findings were observed, indicating that brief exposure to distinct populations had no effect on dimensional mappings. The limitations of the experiment's design were discussed, as was the difficulty of measuring *conceptual* instead of *perceptual* after-effects.

In contrast to Chapters 3 and 4, Chapter 5 examined the representation of another's cognitive process (memory) as opposed to their social traits. It highlighted how the topics of metacognition and theory of mind have tended to be studied separately, focusing respectively on cognitive processes and mental state representations, but that both - and importantly the relationship between them - can be accounted for in the Mind-space framework. Moreover it describes how metacognitive accuracy of one's own mind influences whether the self is a good model for another person's mind and mental states. The experiment investigated how one's own memory, and metacognitive accuracy of such, affected the ability to represent another's memory performance on a dual task with varying levels of difficulty. There was no consistent pattern of results from which to draw clear interpretations. The limitations of the design were discussed, and in particular it was suggested that future studies of representing cognitive processes in Mind-space would benefit from population values from which to derive accuracy scores, in a similar manner to the personality experiments in Chapter 3;

such population values could be drawn from the standardised scores for the cognitive subtasks in Intelligence Quotient measures.

Chapter 6 presented three experiments that were designed to be capable of providing positive evidence of implicit mentalizing, and to examine individual differences therein. Visibility was manipulated in a visual perspective-taking task using two different methods. In Experiment 1, the same reaction time effect was observed for both a social and non-social stimulus, and crucially in both visible and invisible conditions. Experiment 2 showed the reaction time effect in three different visibility conditions. Collectively these experiments provide no support for the hypothesis that implicit mentalizing underpins the effect in this task. Experiment 3 tested whether the 'submentalizing' domain-general process underlying the task effect could be moderated by explicit representation of the human stimulus's mental state, but the findings did not support this hypothesis. Finally, there was no evidence that the reaction time effect in this task correlated with individual differences in perspective-taking and in empathetic tendencies.

7.2 Limitations and Future Directions

7.2.1 Where does Mind-space fit as a theory in the literature?

Mind-space complements rather than replaces or negates existing theories of mental state inference and the dimensional representation of minds. This is attributable to its aim to capture individual differences in mind and mental state representation, which has been an unsolved puzzle that this thesis attempted to address. In reference to Simulation Theory, as discussed in Sections 2.8 and 5.1, Mind-space can explain how the self's location in space can affect the accuracy of representations of others, and indicate when the self is a good model for other minds. With respect to Theory Theory, it still is the case that mental states may be reasoned about using a naïve psychology,

and such reasoning may extend to the Mind-space framework. Mind-space is best accounted for by Godfrey-Smith's (2005) sense of a model of other minds in two ways: first, as a scientific theory it adopts a model-building approach of developing a hypothetical structure (Mind-space) that is used to understand a specific problem (variance in theory of mind); second, as a cognitive model of other minds it accommodates that different target minds will be modelled differently, and that accuracy of mental state inference will depend on an individual's skill in model use.

One criticism made of previous attempts to explain individual differences in theory of mind is that they account for only mental state concept acquisition and domain-general abilities. One might argue that this may be all there is to explain: understanding mental state *content* and the ability to perform the cognitive *computation*. Regarding the latter, Mind-space does not explain variance in processes involved in theory of mind such as decoupling, recursion, or causal inference (Schaafsma et al., 2015). Relatedly, it is not claimed that Mind-space relies on dedicated neural mechanisms, and therefore the skill and the effort with which domain-general resources are employed will still explain some individual differences. Mind-space also does not address variance in the speed or ease with which one acquires mental state concepts, such as those described by Wellman's scale (Wellman & Liu, 2004). The specificity of Mind-space lies in its domain-specific representational structure which maps specific mental states to the mind that generates them (See Figure 7.1). The computations underpinning these inferences are likely to require input from a wide and varying array of cognitive processes, from emotion perception to probabilistic learning. The contribution Mind-space makes is that it (1) links and (2) goes beyond content and computation by presenting a model that can explain and test why individuals differ in what mental states they ascribe to others, or indeed themselves.

One literature in which the Mind-space theory could contribute significantly is Artificial Intelligence (AI). The focus on theory of mind as a uniquely human and specialised, even innate, module has limited the understanding of the mechanisms involved that could contribute to AI implementation (Heyes, 2018); therefore the lack of evidence in Chapter 6 of an implicit theory of mind ability in humans is encouraging for the AI field. Moreover, neuroimaging studies (e.g. Richardson et al., 2018) have done much to locate theory of mind brain activation, but this is not useful for explaining the psychological mechanisms involved (Heyes, 2018). Mind-space, as described in Figure 7.1, provides an implementable framework for an artificial theory of mind. Indeed, Rabinowitz et al (2018) has developed a *Machine Theory of Mind* in which an artificial agent can ascribe false beliefs about the location of an object to other agents with different sight capabilities. This is conceptually similar to the manipulations of agent's paranoia in Chapter 3 Experiments 3 and 4 and its impact on false belief ascription.

Within AI, theory of mind is sometimes described as a meta-learning problem (Rabinowitz et al., 2018). The application of the science of learning could yield much understanding of Mind-space both in humans and in artificial technologies. For example, it has been proposed that theory of mind may be a culturally taught skill (Heyes & Frith, 2014). A subsequent question, and possible source of variance, is knowing whom to learn from and when to learn from others (Heyes, 2016). The use of metacognitive social learning strategies (MSLS), of 'knowing who knows' (Heyes, 2016), may play a significant role in the accuracy of an individual's Mind-space and of the inferences about others' mental states. For example, if an individual has learned an explicit rule about minds that *someone who is impartial is likely to be more accurate*, this may serve as a MSLS resulting in more accurate mental state inferences. Moreover, if this strategy is culturally taught rather than acquired via the trials and errors of personal experience, then this can speed the development of an individual's Mind-

space. Exploring the interaction between MSLS and theory of mind may have important implications. Theory of mind is often thought to have evolved 'for' the prediction of behaviour and understanding communication; The MSLS perspective highlights a quite different, epistemic function (Heyes, *personal communication*, April 2017).

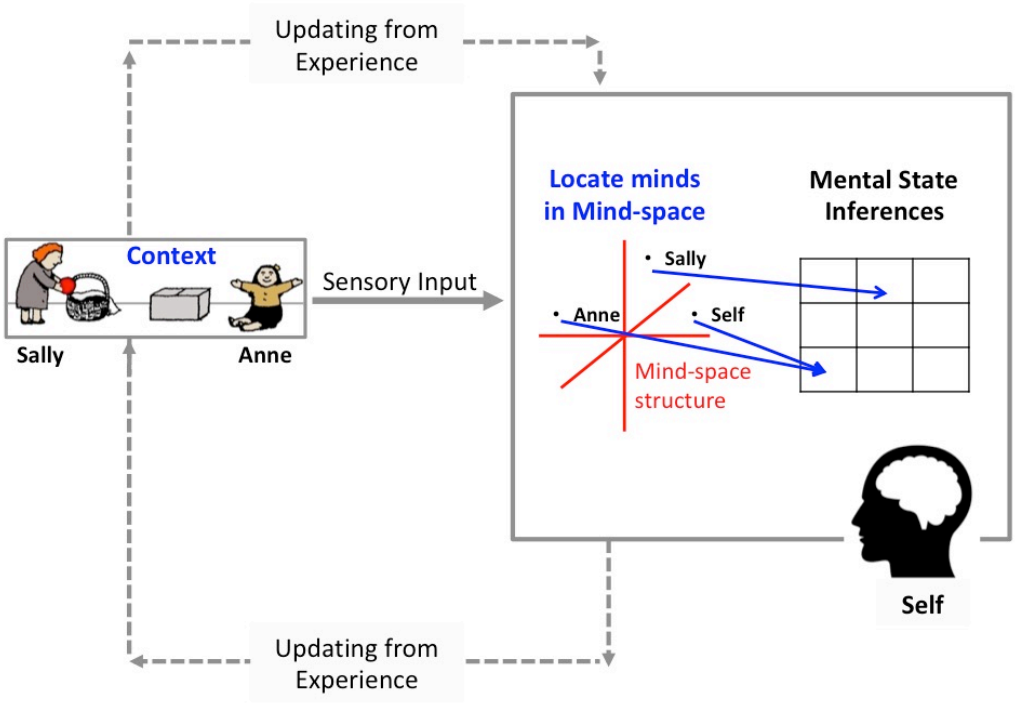

Figure 7.1 An illustrated example of the proposed Mind-space framework. People and contextual factors are perceived via the senses. Minds are then represented along dimensions in Mind-Space. The dimensions can reflect any discriminable aspect of minds. The structure of Mind-space (red axes) reflects the covariance between dimensions, and the granularity of the dimensions. Individual minds are then located in space (e.g. Sally, Anne, and Self). Mental state inferences made for each individual mind depend on a combination of their location within Mind-Space and contextual factors. All aspects of the model are updated through experience; this means that the Mind-space structure, and the mappings between location in space and mental state inferences (blue arrows) will depend on an individual's experience and environment.

7.2.2 Are minds represented in a multidimensional space?

One element of the Mind-space framework is that minds are represented in a multidimensional space (i.e. Mind-space) whereas mental states are represented in a discrete space (See Figure 7.1). This distinction arises from the definitions employed,

whereby minds comprise all cognitive systems (including the enduring social traits exhibited by such a system) whereas mental states reflect the representational content of these systems. This complements one key aspect of Mind-space: accounting for within-individual variance, i.e. that the same mind will have many different mental states and these are predicted by its relatively stable location in space. Furthermore, the operationalization of 'mind' and 'mental state' has by design mapped onto quantifiable characteristics and discrete propositional attitudes respectively.

The dimensional representation of minds accords with previous theories, such as Fiske's warmth vs. competence model which has shown that social categories do map onto a 2-dimensional space (Fiske, Cuddy, & Glick, 2007). The discrete representation of mental states contrasts with Tamir's work (Tamir et al., 2016) which describes 'mental states' as on dimensions, however their definition of mental states corresponds to states of mind, that can be represented on a continuum (e.g. drunkenness, consciousness), rather than propositional attitudes.

To measure mental state inference in Experiments 3 and 4, two discrete mental states were presented (True vs. False Belief Location) and the relative probability was measured. It is possible that individuals represented both mental states (True vs. False Belief) in the agent and the subsequent inference of the 'correct' mental state was based on the relative probability. I suggest that representation refers to the set of all mental states available to an individual (e.g. the grid in Figure 7.1) and inference refers to the reasoning underpinning the selection of one specific mental state (e.g. the specific box within the grid in Figure 7.1). The size of the set and inferential processes involved in selection for a particular target mind is a likely source of individual differences for future studies to explore. This is also one way in which to address the frame problem (Section 1.4.2).

In Chapter 3 Experiments 3 and 4 found supportive evidence for the claim that minds are represented in a multidimensional space, but Chapter 4's experiment did not and Chapter 5's was not optimally designed to test this hypothesis. The distinction between how spaces are described as scientific theories and the extent they actually reflect how the mind and brain are processing information is important to note. For instance, while Face-space has proved useful for advancing the understanding and study of face processing, Burton and Vokey (1998) have demonstrated that psychological claims made about interpreting distance in Euclidian space require certain mathematical assumptions about the dimensional distributions to be held. Based on the theoretical development and empirical evidence presented in this thesis, the current construal (Godfrey-Smith, 2005) of Mind-space as a scientific model is that it serves as a predictive tool for generating testable hypotheses and the dimensional/discrete assumptions may not be met in future work. However the space is described, variance in minds explaining variance in mental states is a key proposal for the understanding of individual differences in theory of mind.

An ambitious new aim for future studies to explore is whether, as a multidimensional space, Mind-space is represented in the brain using cognitive maps (Behrens et al., 2018; Epstein, Patai, Julian, & Spiers, 2017). It has been shown that grid cells, firing in a hexagonal pattern, represent a mental map of spatial relationships in the physical environment (Hafting, Fyhn, Molden, Moser, & Moser, 2005) and the relations between non-spatial concepts (Constantinescu, O'Reilly, & Behrens, 2017; Garvert, Dolan, & Behrens, 2017). Crucially, evidence of cognitive maps has been observed for conceptual stimuli that relate to one another in a dimensional space (Constantinescu et al., 2017) and in a discrete (Garvert et al., 2017) space, but in the latter case the findings did not support the space being Euclidian.

7.2.3 Is Mind-space more than trait-space?

The Mind-space framework defines minds as the entire set of cognitive systems, which includes both cognitive processes and social traits. The idea of a multidimensional 'trait space' in which facial expressions are represented is not new (Stolier, Hehman, & Freeman, 2018; Todorov, Said, Engell, & Oosterhof, 2008); the novel aspect of Mind-space is the mapping of location to mental states given a particular context (Figure 7.1). Section 2.6 presented previous findings that suggest that others' cognitive processes are indeed represented and influence behaviour towards others. While Experiments 3 and 4 in Chapter 3 presented good evidence of the relationship between the representation of minds on social trait dimensions and mental state inferences, Chapter 5's experiment did not provide evidence of the representation of another's memory. The limitations of the design in Chapter 5 were discussed, and it particularly lacked a direct test of the mapping between location on a memory dimension and mental state inference. This thesis therefore provides evidence for the representation of social traits within Mind-space, but not cognitive processes. Whether cognitive processes are represented within Mind-space is therefore a question future studies should address. To note is the potential role of metacognition of cognitive processes for representing the self and other in Mind-space. It is also of interest to determine whether cognitive processes are represented as such, or whether they are instead represented as traits? For instance when inferring another's mental state the content of which depends on remembering something, when might an individual use a trait-like intelligence dimension rather than representing directly the relevant cognitive process (i.e. memory)?

7.2.4 The Double Empathy Problem and Autism

The implications of Mind-space for atypical development were discussed in Section 2.11 however here one additional point is made for the case of autism. This thesis has criticised the emphasis on measuring theory of mind as summing the 'correct' mental state inferences (Section 2.4), and has argued that although 'correct' may be logical or rational, it is confounded with what is typical or the norm. Milton (2012) describes the 'double empathy problem' as a lack of understanding between two people of different dispositional and life experiences. Double empathy stresses the bidirectional nature of social understanding and its mutual and reciprocal properties. Autism is increasingly being appreciated as a source of neurodiversity rather than impairment (Fletcher-Watson & Happé, 2019). In accounting for variance, the Mind-space framework may help with understanding difficulties autistics may have with mental states, and also explain why neurotypicals struggle to understand the minds of those with autism (Sasson et al., 2017; Brewer et al., 2016; Edey et al., 2016). Future studies may explore the possibility of a 'meta-mind-space', in which an individual represents not only their own Mind-space but also others' Mind-spaces, in all their forms.

7.2.5 The case of 'implicit' mentalizing

Chapter 6 has contributed to the on-going debate as to whether certain task effects reflect an implicit ability to represent others' mental states or rather reflect domain-general submentalizing processes (Heyes, 2014). There are considerable multi-lab collaborations currently addressing this question with Open Science practices (Many Babies 2: Infant Theory of Mind, https://osf.io/jmuvd/wiki/home/), and recent studies have failed to find reliably effects that have been used to support the implicit mentalizing account (Kulke et al., 2019; Dörrenberg et al., 2018; Kulke et al., 2018; Kulke & Rakoczy, 2018; Powell et al., 2018; Kulke et al., 2017). However, not finding evidence of implicit mentalizing does not provide support for the submentalizing

account. The experiments in Chapter 6 included a manipulation that specified a pattern of predicted results for the competing hypotheses.

Evidence of implicit mentalizing has been considered of significant theoretical importance for understanding cognitive evolution both in humans and other species (Krupenye et al., 2016), and for explaining theory of mind difficulties in autism (Senju, Southgate, White, & Frith, 2009). However, the lesser-studied alternative of submentalizing has itself many potential implications yet to be fully realised, such as in Artificial Intelligence or in the benefits of a culturally malleable theory of mind for adapting to environmental or societal changes.

7.3 Concluding Remarks

This thesis has presented a new theoretical framework for the study of individual differences in the representation of minds and their mental states. It has presented evidence for the Mind-space framework: that individual differences in the representation of other minds, and in mental state inferences, are attributable to the structure of the space, the ability to locate a target mind in the space, and the mappings between location in space and mental state probabilities. It is a beginning in addressing the problem of conceptualizing variance in theory of mind.

7.4 References

Behrens, T. E. J., Muller, T. H., Whittington, J. C. R., Mark, S., Baram, A. B., Stachenfeld, K. L., & Kurth-Nelson, Z. (2018). What Is a Cognitive Map? Organizing Knowledge for Flexible Behavior. *Neuron*, *100*(2), 490–509. https://doi.org/10.1016/j.neuron.2018.10.002

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

Burton, A. M., & Vokey, J. R. (1998). The Face-Space Typicality Paradox: Understanding the Face - Space Metaphor. *Quarterly Journal of Experimental Psychology Section A: Human Experimental Psychology*, *51*(3), 475–483. https://doi.org/10.1080/713755768

Constantinescu, A. O., Reilly, J. X. O., & Behrens, T. E. J. (2017). Organizing Conceptual Knowledge in Humans with a Grid-like Code. *Science*, *352*(6292), 1464–1468. https://doi.org/10.1126/science.aaf0941.Organizing

Dörrenberg, S., Rakoczy, H., & Liszkowski, U. (2018). How (not) to measure infant Theory of Mind: Testing the replicability and validity of four non-verbal measures. *Cognitive Development*, *46*, 12-30. https://doi.org/10.1016/j.cogdev.2018.01.001

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

Epstein, R. A., Patai, E. Z., Julian, J. B., & Spiers, H. J. (2017). The cognitive map in humans: Spatial navigation and beyond. *Nature Neuroscience*, *20*(11), 1504–1513. https://doi.org/10.1038/nn.4656

Fiske, S. T., Cuddy, A. J. C., & Glick, P. (2007). Universal dimensions of social cognition: warmth and competence. *Trends in Cognitive Sciences*, *11*(2), 77–83. https://doi.org/10.1016/j.tics.2006.11.005

Fletcher-Watson, S. & Happé, F. (2019) *Autism: A New Introduction to Psychological*

Theory and Current Debate (2nd ed.). London: Routledge

Garvert, M. M., Dolan, R. J., & Behrens, T. E. J. (2017). A map of abstract relational knowledge in the human hippocampal–entorhinal cortex. *ELife*, *6*, e17086. https://doi.org/10.7554/eLife.17086

Godfrey-Smith, P. (2005). Folk psychology as a model. *Philosophers' Imprint*, *5*(6), 1–16. Retrieved from https://www.petergodfreysmith.com/FPasModel-PGS-Imprint.pdf

Hafting, T., Fyhn, M., Molden, S., Moser, M. B., & Moser, E. I. (2005). Microstructure of a spatial map in the entorhinal cortex. *Nature*, *436*(7052), 801–806. https://doi.org/10.1038/nature03721

Heyes, C. (2014). Submentalizing: I Am Not Really Reading Your Mind. *Perspectives on Psychological Science*, *9*(2), 131–143. https://doi.org/10.1177/1745691613518076

Heyes, C. M., & Frith, C. D. (2014). The cultural evolution of mind reading. *Science*, *344*(6190), 1243091. https://doi.org/10.1126/science.1243091

Heyes, C. (2016). Who Knows ? Metacognitive Social Learning Strategies. *Trends in Cognitive Sciences*, *20*(3), 204–213. https://doi.org/10.1016/j.tics.2015.12.007

Heyes, C. (2018). Empathy is not in our genes. *Neuroscience and Biobehavioral Reviews*, *95*(November), 499–507. https://doi.org/10.1016/j.neubiorev.2018.11.001

Krupenye, C., Kano, F., Hirata, S., Call, J., & Tomasello, M. (2016). Great apes anticipate that other individuals will act according to false beliefs. *Science*, *354*(6308), 110–114. https://doi.org/10.1126/science.aaf8110

Kulke, L., Johannsen, J., & Rakoczy, H. (2019). Why can some implicit Theory of Mind tasks be replicated and others cannot? A test of mentalizing versus submentalizing accounts. *PLoS ONE*, *14*(3). https://doi.org/10.1371/journal.pone.0213772

Kulke, L., & Rakoczy, H. (2018). Implicit Theory of Mind – An overview of current replications and non-replications. *Data in Brief*, *16*, 101–104. https://doi.org/10.1016/j.dib.2017.11.016

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*, *46*, 97-111. https://doi.org/10.1016/j.cogdev.2017.09.001

Kulke, L., von Duhn, B., Schneider, D., & Rakoczy, H. (2018). Is Implicit Theory of Mind a Real and Robust Phenomenon? Results From a Systematic Replication Study. *Psychological Science*, *29*(6), 888–900. https://doi.org/10.1177/0956797617747090

Milton, D. (2012). On the ontological status of autism: the 'double empathy problem'. *Disability & Society*, *27*(6), 883–887. https://doi.org/10.1080/09687599.2012.710008

Powell, L. J., Hobbs, K., Bardis, A., Carey, S., & Saxe, R. (2018). Replications of implicit theory of mind tasks with varying representational demands. *Cognitive Development*, *46*(September 2017), 40–50. https://doi.org/10.1016/j.cogdev.2017.10.004

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

Richardson, H., Lisandrelli, G., Riobueno-Naylor, A., & Saxe, R. (2018). Development of the social brain from age three to twelve years. *Nature Communications*, *9*(1), 1–12. https://doi.org/10.1038/s41467-018-03399-2

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

Schaafsma, S. M., Pfaff, D. W., Spunt, R. P., & Adolphs, R. (2015). Deconstructing and reconstructing theory of mind. *Trends in Cognitive Sciences*, *19*(2), 65–72. https://doi.org/10.1016/j.tics.2014.11.007

Senju, A., Southgate, V., White, S., & Frith, U. (2009). Mindblind eyes: an absence of spontaneous theory of mind in Asperger syndrome. *Science (New York, N.Y.)*, *325*(5942), 883–885. https://doi.org/10.1126/science.1176170

Stolier, R. M., Hehman, E., & Freeman, J. B. (2018). A Dynamic Structure of Social Trait Space. *Trends in Cognitive Sciences*, *22*(3), 197–200. https://doi.org/10.1016/j.tics.2017.12.003

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. https://doi.org/10.1073/pnas.1511905112

Todorov, A., Said, C. P., Engell, A. D., & Oosterhof, N. N. (2008). Understanding evaluation of faces on social dimensions. *Trends in Cognitive Sciences*, *12*(12), 455–460. https://doi.org/10.1016/j.tics.2008.10.001

Wellman, H. M., & Liu, D. (2004). Scaling of Theory-of-Mind Tasks. *Child Development*, *75*(2), 523–541. https://doi.org/10.1111/j.1467-8624.2004.00691.x