Rethinking Theory of Mind Measurement in Neurotypical Adults:

Response Generation and Selection

by

Kit Ling Yeung

A thesis submitted to the University of Birmingham for the degree of

DOCTOR OF PHILOSOPHY

School of Psychology

College of Life and Environmental Sciences

University of Birmingham

July 2024

Abstract

The research on mindreading in adults has proliferated over the past two decades, but not all measures are equally suitable for assessing individual differences in mindreading performance among neurotypical adults. This thesis presents a systematic review of the measures used to evaluate mindreading in neurotypical adults, identifying measurement challenges and a limited evidence base for assessing the psychometric properties of even popular measures. Additionally, focusing on tasks that present social stimuli without a known ground truth of the mental states of portrayed characters, this thesis proposes alignment, or social agreement, as a practical alternative criterion for mindreading success instead of "accuracy". A series of eight empirical studies were conducted to examine the presence of multiple legitimate mental state interpretations, task-related factors influencing these interpretations, and the role of context in the generation and selection of mental state interpretations. The results challenged the notion of a single best mental state interpretation of ambiguous social stimuli, revealing multiple popular interpretations among participants that varied between groups, and that the format of the task and contextual information about the depicted social interactions influenced mental state interpretations. The findings also provided support for studying the generation and selection of mental state interpretations as distinct processes, with context strongly influencing the selection of the best interpretation while more weakly constraining the generation of plausible interpretations. Possible indices of individual differences in adult mindreading were explored, showing that the tendency to generate multiple interpretations was a more promising direction than alignment and flexibility to adjust interpretations with reference to changes in context. The concluding chapter summarises the findings, discusses the implications and limitations of the current studies, and suggests future research directions for measuring mindreading in neurotypical adults and unravelling the cognitive basis of mindreading.

Acknowledgements

I would like to express my sincere gratitude to my supervisors, Professor Ian Apperly and Dr. Rory Devine, who have provided me with immense support throughout my PhD journey. Their expertise and generous feedback have granted me invaluable insights, and their encouragement allowed me to think critically and independently. They were always responsive to my queries and keen to provide feedback, even during busy times. It has been an honour to work with them, as they are not only exceptional supervisors but also truly thoughtful and kind individuals. This PhD thesis would not have been possible without their guidance.

I also wish to thank everyone in my lab who has been supportive and helpful over the years. Special thanks to Christina for her assistance with trying out the coding scheme for Study 2 and for serving as the second coder for Study 3. I am grateful to Daniel, Sanne, and Rob for the inspiring intellectual exchanges during lab meetings, lunches, and conferences. Thanks to Iris for the casual chats and ongoing food-sharing invitations, which have kept me sane amidst the overwhelming workload. I wish everyone the best in their future endeavours, both professionally and personally. I am also grateful to all my friends who have stayed connected with me, an extreme introvert, despite the distance over the years.

My deepest gratitude goes to my parents for their unwavering support and granting me the autonomy to pursue my PhD studies abroad. Thanks to my brother for always keeping in touch and making me feel connected to the family.

I extend my thanks to the reviewers who peer-reviewed the systematic review reported in Chapter 2. Their insightful comments helped refine the manuscript to a publishable standard in a top journal and encouraged me to think more deeply about the research questions. Lastly, I am profoundly grateful to all the participants who took part in the experiments. This research would not have been possible without their support.

Table of Contents

Chapter 1

General introduction

1.1 Overview

Theory of mind (ToM), also known as mindreading or mentalising, is defined as the ability to represent, reason about, and predict and explain behaviour with mental states. It is a crucial ability for navigating social interactions (e.g., Brüne, 2005; Happé & Frith, 1996; Paal & Bereczkei, 2007; Watson et al., 1999). While early research in mindreading centred around development of this ability, observations of variability in adults' daily mindreading behaviour indicate a research gap in understanding individual differences in neurotypical adults. The first challenge for researchers is to determine how to measure these individual differences. There are numerous ways to measure mindreading, each differing in test format and suitability for testing adults. Theories also suggest that mindreading involves multiple distinct processes, implying that assessments should consider which processes are being captured.

The overarching aims of the current thesis are to address the challenges of measuring and characterising individual differences in mindreading in "neurotypical" adults, as well as exploring the processes of generating and selecting mental state interpretations. The existing measures will be reviewed in Chapter 2. Then, with a focus on mental state interpretation tasks featuring stimuli that do not have a known ground truth, this thesis will discuss the measurement challenge of how "accuracy" should be defined in Chapter 3. The thesis will then examine the individual- and task-related factors influencing interpretations of a target's mental states in Chapters 3-5. The thesis will also explore whether consistent individual differences are observed in the processes of generation and selection of mental state interpretations in Chapters 3-6.

To set up the present thesis within existing empirical findings and theoretical frameworks, the general introduction will provide an overview of (1) the challenges in defining and measuring mindreading in adults, (2) the nature of individual differences in mindreading in adults, and (3) debates on what characterises mindreading success. For consistency, the term “mindreading” will be used most of the time. This decision is based on that the term “theory of mind” presupposes a process of attributing mental states by theorising with a set of concepts and principles, while the term “mentalising”, particularly in literature on mentalisation-based therapy, often refers to a process more akin to mind perception or mind-mindedness rather than the attribution of mental states, as described in Chapter 2. In the current thesis, the study of mindreading focusses on the content of mental states attributed to others, specifically the interpretation of what others are thinking or feeling.

1.2 Challenges in defining and measuring mindreading in adults

1.2.1 Challenges in defining mindreading

Early research on mindreading focused on the acquisition of mental state concepts such as desires, intentions and beliefs in early childhood (e.g., Gopnik & Astington, 1988; Perner et al., 1987; Wimmer & Perner, 1983). According to constructivist accounts, these concepts develop progressively, and children’s possession of a theory of mind is marked by acquiring all the necessary concepts, with the concept of false belief being a benchmark (Gopnik & Wellman, 1992; Perner, 1991). The false belief task captures whether a child understands that individuals act according to what they believe about the world rather than actual states of reality by constructing situations in which the target agent has a belief that does not match reality (Wimmer & Perner, 1983). In later research, five stages of mental state concept development were identified (Wellman & Liu, 2004). In North American and Australian studies, young children progressed from understanding that other people can have different desires, then different beliefs, different knowledge, mistaken beliefs, and feel one emotion but show another (Wellman & Liu, 2004). In Chinese and Iranian children, the ability to distinguish between what two agents know appeared to emerge before the ability to recognise that two people can have different beliefs about the same reality (Wellman et al., 2011, 2018). The sequence of progression was also evident in deaf children, despite showing a delay in development (Wellman et al., 2011, 2018). The discovery of this 5-step Theory of Mind Scale shows that there is an additional stage of hidden emotion after the acquisition of false belief in the development of mental state concepts, but even so, the vast majority of children will have acquired these concepts by middle childhood, with 79% children aged between 7.5 and 11.5 passing this final stage of real-apparent emotion understanding (Peterson et al., 2012; Wellman & Liu, 2004). Despite having acquired the relevant mental state concepts, children in middle childhood and adolescents exhibited variation in mindreading performance that correlated with real-life outcomes (Devine, 2021; Devine et al., 2016; Hughes & Devine, 2015), which indicates that variation in mindreading cannot be fully explained by the acquisition of concepts. Hence, the assessment of mindreading in middle childhood and beyond called for a new way of operationalising mindreading performance, and these measures are called advanced mindreading measures (Osterhaus & Bosacki, 2022).

In contrast to assessments of specific mental state concepts, which are clear about what is measured, the advanced measures capture more complex and less well-defined constructs pertinent to mindreading. There are various operationalisations of mindreading in these measures and the measures have not been found to be consistently interrelated (Apperly, 2010; Happé et al., 2017; Osterhaus & Bosacki, 2022; Schaafsma et al., 2015; Warnell & Redcay, 2019). In the study by Warnell and Redcay (2019) that investigated the interrelations among various mindreading measures in participants including children aged 4 to 12 and adults from an undergraduate sample, the authors failed to identify a unified latent factor underlying the measures included in the study. To account for such findings, apart from criticisms on convergent validity of mindreading measures, which will be explained in the next subsection, mindreading is also argued to be a multidimensional construct that involve multiple subconstructs. There are also researchers who argue some of the tasks capture constructs other than mindreading, such as emotion recognition and anthropomorphism (e.g. Oakley et al., 2016; Tahiroglu & Taylor, 2019; Waytz et al., 2010).

Some theoretical principles are proposed for determining if a task captures the essence of mindreading, such as the necessity of representing mental states and distinguishing the mental states between that of oneself and others (Quesque & Rossetti, 2020). However, tasks meeting these criteria might only assess a specific aspect (i.e., self-other distinction), within the broader construct of mindreading which arguably also involves motivational elements and other abilities (e.g., Apperly, 2012). In this thesis, a general definition of mindreading is adopted: it is characterised as the interpretation of others' mental states.

Nevertheless, regardless of whether a broad or narrow definition is used, researchers face similar challenges in measuring individual differences in mindreading in adults, as most existing measures were not designed for this purpose and surprisingly few studies have assessed the psychometric properties of these measures. While it is possible that some of these measures may still be effective as measures of individual differences in mindreading, it is unlikely that all existing measures are equally suitable for assessing mindreading in neurotypical adults or demonstrate satisfactory psychometric properties.

1.2.2 Challenges in measuring mindreading

There are two major challenges when measuring mindreading in adults. First, not all measures are sensitive to variance in performance within neurotypical adults, as many tasks were designed for children or for comparing clinical and neurotypical populations. Ceiling effects are likely to be observed in the tasks designed for children as they were designed to capture the progression of mental state concept understanding and theoretically, adults are presumed to possess all these concepts (Peterson et al., 2012; Wellman & Liu, 2004). The mental state concept account for mindreading posits that all such variations are measurement errors, which is unlikely true, as mindreading performance in early and middle childhood exhibits rank-order stability over time and correlates with real-world social outcomes such as social competence (e.g., Devine, 2021; Devine et al., 2016; Hughes & Devine, 2015). Whether the measures used to test older children and adolescents are still sensitive to variation in adults and whether these meaningful individual differences persist into adulthood warrants further research. However, the above findings indicate that the application of methods that focus on detecting developmental differences to assess individual differences in adults (e.g., El Haj et al., 2017) should be viewed with some caution, as we can expect that they are likely to mask any variation in adults' mindreading performance due to ceiling effects. A similar problem in intelligence testing has long been identified: it is problematic to assume no individual differences in intelligence in adults when they are tested with items devised for children, and vice versa (Anastasi, 1948).

An analogous problem exists for tasks designed to detect differences between experimental conditions or between clinical and non-clinical groups, as a well-designed task for comparing between different conditions aims to minimise between-participant variation to maximise sensitivity to detect between-condition differences (Hedge et al., 2018). Similarly, a task designed to be sensitive for detecting differences between clinical and neurotypical populations is also unlikely to be optimised for detecting individual differences within the neurotypical group.

The second measurement challenge pertains to the psychometric properties of the existing tasks, including reliability and validity. Reliability and validity are established in multiple facets and it is important to assess the psychometric properties of a task to determine whether it produces consistent results and measures the construct of interest; while the former concerns the concept of reliability, the latter pertains to validity (Rust et al., 2020).

Reliability is the extent to which repeated measures correlate, thus capturing variance other than measurement error. It can be approximated by taking repeated measurements, according to the classical test theory (Rust et al., 2020). If test items in a measure capture the same construct, they should correlate and show good internal consistency (Fu et al., 2023; Revelle & Condon, 2019), even if items present different contexts or settings, or have different levels of difficulty (Devine & Hughes, 2016). Test-retest reliability should also be demonstrated in terms of stable mindreading performance over short periods (Rust et al., 2020), if mindreading is a trait-like ability (e.g., Devine, 2021). Additionally, inter-rater reliability should be examined for tasks scored from open-ended responses to ensure consistent scoring (Devine, Kovatchev, Grumley Traynor, Smith & Lee, 2023).

Validity is upheld if a test measures the intended construct; it is a matter of degree (Nunally, 1978), and evidence of validity is assessed in multiple ways. Criterion-related validity is assessed by investigating how well the measure predicts relevant but distinct variables (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 2014). Hence, associations between task performance and behaviour, traits, or outcomes with real-life implications, are expected, especially social outcomes for the case of mindreading (e.g., Banerjee et al., 2011; Canty et al., 2017; Imuta et al., 2016; Devine et al., 2023). Convergent validity involves examining associations between similar or identical constructs, in other words, mindreading measures adopting different stimuli or formats, as well as other measures of social cognition, should be associated with one another. It is also imperative to examine the discriminant validity of a measure as it ensures the measure does not capture unintended constructs (Rönkkö & Cho, 2022).

It is important to review the literature for evidence supporting the psychometric properties of existing measures to justify their use. In the existing literature, review articles evaluate the psychometric properties of mindreading measures used in children (Beaudoin et al., 2020; Fu et al., 2023; Ziatabar Admadi et al., 2015), measures assessing alexithymia alongside mindreading (Pisani et al., 2021), or do not specifically examine the psychometric properties of measures for adults (Derksen et al., 2018; Osterhaus & Bosacki, 2022). In contrast, this thesis will present a review focused on measuring mindreading in adults, which reveals a limited evidence base to support the effectiveness of existing measures in assessing individual differences in mindreading among neurotypical adults.

1.3 The nature of individual differences in mindreading in adults

The previous section described the measurement challenges in capturing individual differences in mindreading with an assumption that such individual differences do exist. The existence of genuine individual differences in mindreading has been systematically scrutinised and supported by research on children, which has shown variation in mindreading performance exhibits rank-order stability over time and is associated with real-life outcomes, indicating the individual differences in mindreading in children are enduring and meaningful (Dunn et al., 1991; Devine, 2021). Such studies in adults are sparse, but the assumption of the existence of such individual differences is aligned with our everyday intuition that individuals vary in their tendency and ability to take the perspectives of others, even then these individuals have similar cognitive abilities. For instance, academics, who are likely to have high levels of cognitive ability, can exhibit considerable variation in mindreading, which indicates a likely unique social component in these differences.

Moreover, studies examining mindreading performance between neurotypical adults varying along broader neurodivergent phenotypes support this idea. From the perspective that the entire population varies in broader phenotypes, neurotypical individuals are contrasted with neurodivergent individuals based on how common their neurocognitive profiles are within the population. People whose neurocognitive profiles are more common are described as “neurotypical”, while “neurodivergent” individuals have profiles that are less frequently observed and diverge from dominant societal standards (Pellicano & den Houting, 2022). It has been shown that non-clinical individuals scoring high in schizotypy tended to perform worse on mindreading tasks compared to those scoring low in schizotypy, in a manner similar to the performance gap between individuals with schizophrenia and neurotypical individuals (Kocsis-Bogar, et al., 2017). These observations suggest the presence of genuine individual differences in mindreading even in neurotypical adults.

While psychometrically sound measurements that are sensitive to variance within the neurotypical population are necessary to detect individual differences in adult mindreading, it is equally important to understand the nature of these differences. As noted in the previous subsection, if the variations are genuine and not merely due to measurement errors, they should correlate with meaningful outcomes such as social functioning. This is consistent with the social individual differences account of mindreading first proposed by Dunn et al. (1991), which posits that enhanced mindreading skills have a significant effect on social functioning beyond other general cognitive abilities (Apperly, 2012); evidence supporting this has been found in research on individual differences in mindreading during early and middle childhood (Devine, 2021; Devine et al., 2016; Hughes & Devine, 2015).

If individual differences in mindreading in adults cannot be attributed solely to mental state concepts, what do these differences signify? Perhaps individuals vary in their ability to use these mental state concepts, especially when it becomes challenging to combine these concepts for decision-making, with reference to research indicating that negative desires and false beliefs are more cognitively demanding even in adults, which suggest complexity in using these concepts (Apperly et al., 2008; 2011). However, these experimental studies generally reveal overall tendencies in adults rather than individual variations in abilities. It is also plausible that people differ in general cognitive abilities, such as executive functioning and language skills. For instance, studies have shown a bilingual advantage in mindreading performance in both children and adults (Bialystok & Senman, 2004; Bialystok &

Viswanathan, 2009; Goetz, 2003; Kovacs, 2009; Diaz & Farrar, 2018, Navarro & Conway, 2021). Some researchers attribute this to bilingual individuals' superior executive functioning (e.g. Bialystok & Senman, 2004), while others suggest it is due to their enhanced metalinguistic awareness (e.g. Kloo & Perner, 2003). Moreover, researchers have proposed various sources for these individual differences in adults, such as the ability to locate a mind within a mind-space (Conway et al., 2020) and the flexibility to make mental inferences based on varying contexts (Devine, 2021; Hughes & Devine, 2015). Nevertheless, before exploring the origins of the differences in mindreading performance, it is important to acknowledge that individuals can differ in mindreading in multiple ways, and these different facets are captured by different measures.

1.3.1 Facets of mindreading measured

Understanding how individuals differ in mindreading involves examining various facets of this construct. The existing literature often categorises mindreading into several dichotomies. This section discusses three such dichotomies: cognitive versus affective mindreading, mental state decoding versus reasoning, and accuracy versus propensity.

Cognitive vs. affective mindreading. Mindreading is often divided into cognitive and affective facets, also known as "cold ToM" (i.e., cognitive mindreading) and "hot ToM" (i.e., affective mindreading) (e.g., Meinhardt-Injac et al., 2020). Neuroscience research indicates distinct neural activation patterns when tasks involve cognitive versus affective mental states (Sebastian et al., 2011; Shamay-Tsoory & Aharon-Peretz, 2007; Poletti et al., 2012). Cognitive mindreading involves understanding the beliefs and intentions of others, with the false belief task being a representative cognitive mindreading task, while affective mindreading requires the understanding of others' emotions (Mitchell & Phillips, 2015). Affective mindreading tasks may, for example, present visual stimuli requiring participants to identify the target’s emotions or present verbal vignettes that involve identifying the emotions of targets.

However, while some tasks clearly distinguish the two types of mindreading or include both as distinct components of the task (e.g., the Yoni task; Shamay-Tsoory & Aharon-Peretz, 2007), the distinction between cognitive and affective mindreading is not always clear-cut in advanced mindreading tests, as both cognitive and affective mental states are found to be intertwined in many tasks, for example, the Faux Pas Recognition task (FPRT; Baron-Cohen et al., 1999) and the Movie for the Assessment of Social Cognition task (MASC; Dziobek et al., 2006). Even for the widely-used Reading the Mind in the Eyes Task (RMET), researchers have claimed that it measures cognitive mindreading (Gregory et al., 2002; Sharp, 2008) and affective mindreading (Richell et al., 2003; Tonks et al., 2007), despite adopting the identical task. Hence, strictly distinguishing cognitive from affective mindreading might be impractical.

Mental state decoding versus reasoning. Mental state decoding tasks, also known as social-perceptual tasks, are contrasted with mental state reasoning tasks, often referred to as social-reasoning or social-cognitive tasks (Meinhardt-Injac et al., 2020). Decoding tasks involve immediate interpretations from observable cues, such as recognising emotions from facial expression. The most representative example is the commonly used RMET (Baron-Cohen et al., 2001), which presents photos of the eye region and requires participants to select one out of four options suggesting the emotion or cognitive state of the target. Decoding tasks predominantly present stimuli visually, but there are also tasks that present auditory stimuli featuring human voices (e.g., Reading the Mind in the Voice task; Golan et al., 2007). In contrast, reasoning tasks require more complex processing, such as deducing an agent’s intentions from a vignette. For example, the Strange Stories Task (Happé, 1994) requires participants to infer the reason behind a target’s speech or action from the story presented.

Some tasks, like those presenting videos of social interactions (e.g., the MASC; Dziobek et al., 2006), involve both decoding and reasoning, as they require interpreting facial expressions and body language while understanding the mental states of characters considering the context in which the social interaction is situated.

Accuracy versus propensity. Accuracy in mindreading is assessed by comparing participants' responses to pre-determined "correct" answers. In contrast, propensity measures the motivation or tendency to attribute mental states to agents without evaluating the appropriateness of these attributions. Propensity is considered a motivational component of mindreading (Carpenter et al., 2016; Contreras-Huerta et al., 2020), distinct from accuracy (Devine & Apperly, 2022; Carpenter et al., 2016), and is commonly measured by questionnaires that focus on capturing trait-like motivation (e.g. Carpenter et al., 2016), although sometimes also in behavioural assessments by methods such as calculating the proportion of mental state attributions in an open-ended response (e.g. Dodell-Feder et al., 2013). Most existing tasks focus on capturing accuracy. There are also tasks that measure accuracy and propensity in different subscales, by considering both the extent to which descriptions are made with reference to mental states and the appropriateness of mental state attributions, respectively (e.g., Animations task; Abell et al., 2000; Castelli et al., 2000)

1.3.2 The processes of generation and selection

The distinction between accuracy and propensity can also be discussed from the perspective of processes involved in mindreading: a possible operationalisation of the propensity to make inferences is the tendency to generate multiple candidate interpretations of a target's mental states, regardless of their appropriateness. This contrasts with accuracy, which implies a selection of the most appropriate interpretation from the possibilities generated.

The concept of selecting from multiple mental state inferences is central to conventional mindreading theories that follow a modular approach, such as the ToMM-SP theory (Leslie et al., 2004), and formal models proposed in more recent research, such as the Bayesian Theory of Mind (BToM) model (Baker et al., 2017). The ToMM-SP theory suggests that mindreading involves a “Theory of Mind Module” that generates possible belief contents and a “Selection Processor” that selects among these beliefs (Leslie et al., 2004), while the BToM model proposes that initial confidence levels or probabilities are assigned to candidate hypotheses for mental state attributions, which are then adjusted in specific contexts to guide the final selection (Baker et al., 2017). Both models agree on a key point: while multiple hypotheses can be generated initially, some are ultimately rejected while the final selection is in favour of the most appropriate interpretation. These theories provide a theoretical basis suggesting the processes of generating and selecting mental state interpretations can be, and should be, studied as distinct phenomena. However, these models primarily focus on simple, highly constrained scenarios involving basic mental state concepts, represented by beliefs and desires. The complexities involved in generating and selecting appropriate mental state interpretations in more naturalistic, contextualised settings remain largely unexplored.

1.4 Debates about and alternatives to “accuracy”

The previous section demonstrated that accuracy and propensity are related but distinct concepts regarding mindreading, each linked to different processes. Most existing tasks focus on assessing mindreading performance based on participants’ accuracy. However, it is controversial whether the “correct” answer truly represents the “accurate” mental state of the target, or the “ground truth.” This controversy largely depends on the design of the tasks. If the correct answer in a task is based on the target agent’s self-reported mental states, it can be considered to capture the ground truth. However, this approach is rare and is only seen in one existing task that claims to measure mindreading, the Interview Task (Long et al., 2022).

The “empathic accuracy” task, which, although not specifically aimed at measuring mindreading, uses a similar approach (Ickes et al., 1986; Ickes, 1993). These two tasks use the target’s self-reported mental states and traits as the basis for establishing ground truth. However, findings that empathic accuracy performance was better explained by expressivity of the target and the participants’ familiarity with the target than trait empathy (Zaki et al., 2008; 2009) cast doubt on whether this type of paradigm is always effective in assessing mindreading ability. Additionally, the reliability of introspective reports of one’s own mental states can be questioned, for example, individuals can fail to notice certain emotions they had due to a lack of attention, or be unable to make a correct judgment of the emotion (Trnka & Smelik, 2020). Furthermore, studies on cognitive dissonance have also shown that individuals’ recall of thoughts or feelings could change when confronted with inconsistencies with their other beliefs (e.g., Festinger, 1957), and research on autobiographical memory has shown that individuals can make mistakes when recalling events about themselves, whether these events happened long ago or relatively recently (Hyman & Loftus, 1998). Despite these concerns, the target’s self-report of mental states provides a reasonable basis for evaluating “accuracy” of a mindreading response. Conversely, when the target agent’s self-report is unavailable, as in the case in almost all existing tasks, whether the “correct” answer is indeed “accurate” becomes questionable.

On one hand, mental state reasoning tasks often use vignettes in various formats such as text, comics, stories, or videos. These vignettes are designed to convey specific mental states as intended by the creator. Therefore, a response matching the “correct” answer may reflect the author’s intention. However, it is debatable whether the author’s intention accurately represents the mental states of the characters, given that these characters do not possess genuine mental states like real people. The notion of “accuracy” is even more problematic in mental state decoding tasks, where the actual mental states of the individuals in the stimuli are often not available. The stimuli of these tasks often did not present targets who were known to be truly experiencing a certain mental state, making direct access to the ground truth impossible. Furthermore, the concept of "decoding" rests upon the assumption that a "correct" mental state can be identified immediately from observable cues (e.g., Harkness et al., 2005; McGlade et al., 2008) that matches the pre-determined answer. However, it is unclear whether the "correct" answers in these tasks are indeed representative of the most appropriate interpretations, as these are often based on the opinion of researchers or, only less frequently, based on pilot studies. Additionally, it is uncertain whether the answer deemed the most appropriate would vary with other factors, undermining the notion of defining mindreading success as successful "decoding". It has been argued that "decoding" should take into consideration that individuals can infer different meanings from the same facial expressions as the situation differs (Bora et al., 2006), but this suggestion has not been incorporated in the commonly used mental state decoding tasks. Given these issues, it may be necessary for researchers to reconsider what constitutes mindreading success when the ground truth is not directly accessible.

1.4.1 Alternative viewpoints on what characterises mindreading success

The question of ground truth regarding mental states has longstanding viewpoints from philosophical theories. Fodor's (1990) realism perspective asserts that there are objective facts about an individual's mental states, and mindreading success is characterised by accurately capturing these facts. This aligns with the assumption in existing tasks that distinguish correct from incorrect mental state interpretations. In contrast, Dennett's (1987) "intentional stance" account suggests that mindreading success is not about capturing factual mental states, which may or may not exist, but about whether the attribution of mental states successfully predicts or explains behaviour. These perspectives differ fundamentally: realism posits that facts exist prior to mental state ascription and are used for verifying the success of the ascription attempt, while the intentional stance holds that the success of an attempt to ascribe mental states is judged by its success in predicting or explaining behaviour, irrespective of an objective ground truth.

The intentional stance provides an alternative to focusing on “accuracy” as identifying facts about mental states. However, this perspective makes it difficult to distinguish good from poor performance in mental state decoding tasks, as these do not involve behaviour prediction. If the goal is to explain facial expressions or body language by attributing mental state interpretations, multiple interpretations are possible, and there is no clear method to differentiate better from worse interpretations.

If interpretations are to be differentiated in terms of appropriateness, a baseline for comparison is necessary, even if an objective criterion for accuracy is absent. One potential criterion is to compare an interpretation with the consensus of a population, characterising successful mindreading by agreement with other members of a group (Apperly et al., 2024). Drawing from studies on social coordination that have successfully quantified an individual’s level of agreement with other individuals (Mehta et al., 1994; Perez-Zapata & Apperly, 2022), the level of social agreement is labelled as “alignment” in the current thesis.

How one is likely to attribute mental states to others is influenced by practical considerations, such as cultural norms and personal experiences. People sharing similar backgrounds have a higher tendency to share similar norms and experiences and thus, are likely to make similar assumptions which underlie their interpretations of others’ mental states (Apperly et al., 2024). Therefore, even when there is no objective way to determine which interpretations are most accurate, there exists a possible way to evaluate how good someone’s interpretations are – by evaluating how well the interpretations align with those of others in a group.

Furthermore, mindreading is hypothesised to play a crucial role in social interaction and communication. Therefore, given the practical importance of social consensus on mental states in navigating social lives, focussing on one's agreement with others on their mindreading decisions is a functional approach to defining mindreading success.

1.4.2 Possible factors influencing group consensus on mental state interpretations

The view that alignment characterises mindreading success implies that consensus within the group may be influenced by the characteristics of the specific group, which starkly contrasts with having a pre-determined "correct" answer assumed to be a proxy of ground truth that applies to any test takers. This implies that measurement invariance should not be assumed, as different groups may reach different consensus.

Individual factor: Group membership. The idea that different groups of people make different mindreading interpretations and hence find difficulty understanding people from a dissimilar group is laid out in the theoretical framework of the double empathy problem, which posits that autistic individuals find it challenging to interpret the mental states of their neurotypical counterparts, and vice versa (Edey et al., 2016). Findings showing differences in how autistic and neurotypical people express emotions corroborate this proposal (Brewer et al., 2016). Additionally, research within neurotypical groups shows that the expression of emotions also differs between cultures (Jack et al., 2012), and people tend to score higher when interpreting the minds of agents from their own culture (Adams et al., 2010; Perez-Zapata et al., 2016) or when they have higher perceived familiarity with the targets (Zaki et al., 2009). These all suggest the assumption of a single correct answer may not apply across different groups of individuals.

Task-related factors: Task format and context. The way in which a task is administered should also be considered when the basis of mindreading success is determined by the opinions of other people. Task format can significantly affect participants' performance, as evidenced by parallel research in memory. Recognition tasks often yield different performance outcomes compared to free-recall tasks (e.g., Jacoby et al., 1979), with memory deficits being less detectable in recognition tasks. Similarly, in social cognition, participants tend to perform better in forced-choice formats than in open-ended formats (Cassels & Birch, 2014). It is suggested that participants can rule out foils and arrive at the correct answer by elimination in a forced-choice task, whereas such strategies are inapplicable in an open-ended task. Moreover, a study by Betz et al. (2019) has found that individuals had a higher tendency to attribute mental states in the forced-choice version of the RMET in comparison to an open-ended version of it, implying measured propensity of mindreading is also influenced when individuals engage in recognition versus generation of possible mental state attributions.

Another task-related factor is the contextual information provided during the task, including background information about when and where the social interaction takes places, and information about the target. When more contextual information about the interaction is available, interpretations may change, but this is often overlooked in many studies of mindreading (Spaulding, 2018). Cognitive and social psychology have a long history studying the effect of context, showing that people organise their knowledge about the world in scripts, schemas, and stereotypes (e.g., Cantor et al., 1982; Gilbert, 1998; Schank & Abelson, 1977), which are intuitive cognitive structures. These cognitive structures help people maintain shared expectations about social interactions, which enables coordinated behaviour in social situations. Such mental frameworks also exist in handling the perception of personality traits of individuals. The mind-space theory posits that the perception of personality traits is organised within a multi-dimensional space, where each trait is represented as a distinct dimension, and correlated traits are depicted as correlated dimensions (Conway et al., 2019; Long et al., 2022). Research supports that the ability to represent these correlations among personality dimensions is associated with an individual's performance in advanced mindreading tasks. Although the cognitive structures mentioned above are generic and do not provide an explicit answer to what a target is thinking or feeling, they are likely to help individuals narrow down the range of possible interpretations of the target's mental states.

In addition to helping individuals perceive social situations with a well-structured framework of generalised knowledge, neuroimaging studies have shown increased brain synchrony when context is provided (Hasson et al., 2012), suggesting a neural basis for how context modulates social interactions. In the realm of mindreading, recent research on mindreading indicated that participants became less likely to choose the default answer in a change-of-location false belief task when provided with more, particularly inconsistent, information about the target (Cho et al., 2022); a similar effect was also observed when participants were informed that the target had a high level of trait-paranoia or that the character who moved the item in the target's absence had a high level of trait-dishonesty (Conway et al., 2019). In mental state decoding, research has shown that the same facial expression can convey different emotions depending on the context (Aviezer et al., 2012), despite the conventional belief that facial expressions are key cues for "universal emotions" like fear and anger. Such findings suggest that interpretation of observable cues is influenced by contextual information, contrasting with the view that decoding from observable cues is direct and can be achieved without context.

Generation and selection in context. As discussed in the previous sections, the generation and selection of mental state interpretations can be conceptually and methodologically studied as distinct processes. It is also evident that context likely plays a role in mental state attributions. However, it remains unclear whether context influences only the selection of the most appropriate interpretation or also affects the initial generation of candidate interpretations. The effect of context on the generation process has been sparsely, if ever, studied in existing literature.

1.4.3 Flexibility in mindreading

Considering the individual-related and task-related factors discussed in the previous section, it is important to recalibrate consensus to fit the specific population and the format of the task, for consensus can vary significantly across different groups and contexts. A mindreader who aligns well within their own group or similar groups may struggle to perform well when required to align with a dissimilar group or in a different context, where the consensus can differ.

Therefore, a skilled mindreader might be characterised by their flexibility in making mental state interpretations across varying contexts (Devine, 2021; Hughes & Devine, 2015). This involves two key processes. The first process is generating multiple possible interpretations, which aligns with the notion that mindreading resembles adaptive reasoning, which involves individuals generating multiple, modifiable hypotheses to explain a social scenario (Hayward et al., 2018). The second process is selecting the most appropriate interpretation based on the current context and the group to be aligned with. To summarise, possible indices of individual differences in mindreading ability, especially in mental state decoding tasks, include both alignment within group, as well as flexibility to adjust interpretations to align with various groups and contexts.

1.5 Overview and scope of present studies

To summarise the discussion above, mindreading has various definitions and there are challenges in measuring mindreading in neurotypical adults. In the empirical chapters of the current thesis, neurotypical adults are practically defined as adult participants who have not been diagnosed with autism spectrum disorder (ASD). While individual with other forms of neurodivergence are not screened out from the samples, individuals with and without ASD have been specifically shown to exhibit different patterns of social performance, as discussed in the context of the double empathy problem (Alkhaldi et al., 2019; Edey et al., 2016; Milton, 2012). Adopting a broad definition of mindreading and focusing on mental state decoding tasks, the notion of "accuracy" is found to be problematic. Instead, skilled mindreaders might be characterised by their alignment within specific groups as well as their flexibility in adjusting mental state interpretations to cater to varying groups and diverse contexts. However, it remains an open question how to measure alignment and flexibility, and whether these are reliable, consistent indices of individual differences in mindreading. Furthermore, mindreading involves two separate processes, generation and selection, that should be studied independently.

With reference to these challenges in measuring and characterising individual differences in mindreading, this thesis aims to review the current state of research on measuring individual differences in adult mindreading and explore possible indices of individual differences in both the generation and selection processes in adult mindreading. The validity of the ground truth assumption in mental state decoding measures that adopt stimuli without a known ground truth is examined, and factors influencing mental state interpretations are explored. With a special emphasis on the effect of context, this thesis studies generation and selection as distinct processes involved in mindreading.

This thesis presents the findings from a systematic review and a series of empirical studies. The aims of each chapter are listed as follows. Chapter 2 aims to review the psychometric properties of the existing measures systematically and determine whether there is sufficient evidence to support the use of these measures in assessing individual differences in mindreading in neurotypical adults. Chapter 3 features three empirical studies that examine the validity of the assumption of ground truth, offer "alignment" as an alternative perspective to the notion of "accuracy", and test the reliability of indicators of individual differences in generation and selection of possible mental state interpretations. The correlation between propensity and alignment is also examined. Chapter 4 further challenges the notion of having a single typical way of decoding mental states when participants are required to engage in the process of recognition contrasted to generation of mental state interpretations. With a focus on the role of context as a task-related factor, Chapter 5 presents three studies to explore the influence of context on the selection of mental state interpretations and to investigate individuals' flexibility to adjust mental state interpretations with varying contexts as another possible index of consistent individual differences. Further grounded on the role of context in mindreading with an attempt to bridge the processes of generation and selection, Chapter 6 features a study that examines if context constrains one's selection and generation of interpretations. Finally, all the studies reported across Chapters 2 to 6 are synthesised in Chapter 7, the general discussion, which wraps up the thesis by presenting the overarching conclusion, limitations, and implications of these studies.

Chapter 2

Measures of individual differences in adult theory of mind: A systematic review

Note: This chapter has been peer-reviewed and published in the journal *Neuroscience and Biobehavioral Reviews*.

Yeung, E. K. L., Apperly, I. A., & Devine, R. T. (2023). Measures of individual differences in adult theory of mind: A systematic review. *Neuroscience & Biobehavioral Reviews*, 105481. https://doi.org/10.1016/j.neubiorev.2023.105481

2.1 Introduction

Theory of mind (ToM), also commonly referred to as mindreading or mentalising, is the ability to represent mental states, reason about them, and make use of them to predict and explain behaviour (Apperly, 2010; Baron-Cohen et al., 1985; Premack & Woodruff, 1978). It is regarded as an important ability that facilitates social interaction (e.g., Brüne, 2005; Happé & Frith, 1996; Paal & Bereczkei, 2007; Watson et al., 1999). Early research on the topic focused on mindreading development in early childhood (e.g., Gopnik & Astington, 1988; Perner et al., 1987; Wimmer & Perner, 1983) and in people with clinical conditions, especially autism (e.g., Baron-Cohen, 1985; Yirmiya et al., 1998; Hughes et al., 2000). There is now clear evidence that mindreading development continues across middle childhood and adolescence (e.g., Apperly et al., 2011; Devine & Hughes, 2013; Hughes, 2016; see Devine, 2021, and Weimer et al., 2021, for a review). Alongside developmental work on children and adolescents, studies of mindreading in neurotypical adults, focused on underlying cognitive and neural processes and the presence of individual differences, have also emerged (e.g., Apperly, 2010; Bradford et al., 2015; Mahy et al., 2014; Qureshi et al., 2020; Schurz et al., 2014). Despite this ongoing interest in mindreading, there is little consensus on how best to measure individual differences in mindreading in neurotypical adults. In this systematic review, we identify two major challenges in measuring individual differences in mindreading performance in neurotypical adults, identify existing measures, and critically examine the measurement characteristics of these measures. The over-arching aim of this review is to take stock of work needed to evaluate existing measures and to develop new ones.

2.1.1 Studying mindreading in adults

Research on mindreading in adults has proliferated in the previous two decades (Apperly, 2021). Neurotypical adults are considered developmentally mature in their understanding of mental state concepts (Apperly et al., 2009; Karmakar & Dogra, 2019), providing a baseline for comparison with other populations such as children and clinical groups. However, adults still show patterns of performance on mindreading tasks that are analogous to those observed in children, such as demonstrating egocentric biases when they need to take the perspective of a less-informed person (Keysar, 2000, 2003), and making inaccurate mental inferences of what another person thinks or feels (Ickes et al., 2000), with notable variation in performance between individuals. From such observations, the study of individual differences in adult mindreading performance has emerged as a meaningful research topic. For example, researchers have suggested various sources of such individual differences, including the ability to locate a mind within a mind-space (Conway et al., 2020), or the flexibility to make mental inferences based on varying contexts (Devine, 2021; Hughes & Devine, 2015). There is also research that teases apart adults' mindreading ability to make accurate mental inferences and their propensity, or motivation, to use their mindreading (Apperly & Wang, 2021; Carpenter et al., 2016; Devine & Apperly, 2022).

Furthermore, researchers have investigated whether adults' mindreading performance correlates with various social skills, cognitive abilities, and traits related to psychiatric and neurodevelopmental conditions (e.g., Abu-Akel et al., 2015; German & Hehman, 2006; McGarry et al., 2021; Nilson & Duong, 2013; Weinstein, Whitemore, & Mills, 2022). Critically, however, the research described above requires that individual differences in mindreading in adults can be reliably and validly measured. There are two problems that should raise concerns about current measures.

Problem 1: Measures may not be sensitive to variance in performance in neurotypical adults. Many studies of individual differences in adults have either employed tasks originally designed for children or for investigating differences between neurotypical adults and adults with psychiatric or neurodevelopmental conditions. According to one account, children acquire an understanding of mental concepts sequentially (Wellman & Liu, 2004). The first concepts include desire, belief and emotion, and subsequent studies suggest that more complex concepts, such as belief-desire reasoning, require the integration of simpler mental state concepts. Empirically, children perform well on all concepts by middle childhood, leaving little possibility of variation in adults. For example, Peterson et al. (2012) found that half of the children aged between 6 to 7.5 passed the hidden emotions task, the most difficult task in the 5-step Theory of Mind Scale (Wellman & Liu, 2004), and 79% children aged between 7.5 and 11.5 were able to pass it. Moreover, the dominant theoretical interpretation considers these findings to chart the acquisition of the concepts that adults are presumed to possess (Peterson et al., 2012; Wellman & Liu, 2004). This interpretation has no capacity to explain variation in the performance of older children and adults, other than as measurement errors in assessing their underlying conceptual competence (Apperly, 2012). If the source of variation in performance on theory-of-mind tasks is indeed measurement error, then individual differences in performance should not be associated with meaningful outcomes (e.g., Hughes & Devine, 2015). However, drawing on research showing that on individual differences in mindreading performance in early and middle childhood exhibit rank-order stability over time and correlate with real-world social outcomes such as social competence (e.g., Devine, 2021; Devine et al., 2016; Hughes & Devine, 2015), it is more likely that these individual differences are meaningful, rather than mere measurement errors. Whether the measures used to test older children and adolescents are still sensitive to variation in adults and whether these meaningful individual differences persist into adulthood warrants further research, the above findings indicate that the application of methods that focus on detecting developmental differences to assess individual differences in adults (e.g., El Haj et al., 2017) should be viewed with some caution, as we can expect that they are likely to mask any variation in adults' mindreading performance.

An analogous problem exists for tasks designed to detect differences between experimental conditions or between clinical and non-clinical groups. A well-designed task for comparing between different experimental conditions aims to minimise between-participant variation to maximise sensitivity to detect between-condition differences (Hedge, Powell, & Sumner, 2018). By extension, a task designed to be sensitive for detecting differences between clinical and neurotypical populations is also unlikely to be optimised for detecting individual differences within groups. Although tasks designed on this basis may still be good measures of individual differences within neurotypical adults this should not be taken for granted.

Problem 2: Psychometric properties of measures. Classical test theory provides a framework for evaluating the quality of measures of psychological constructs such as mindreading (Fu et al., 2023) and has been applied in to evaluate measures of children's mindreading (e.g., Hughes et al., 2000; Devine & Hughes, 2016). According to the classical test theory, a true score on a construct can be approximated by taking repeated measures of it (e.g., Rust et al., 2020). Reliability is characterised by the extent to which repeated measures correlate with one another, as it captures the variance that is not attributed to measurement error of individual tests. Assuming all items in the same measure capture the same construct, the items should correlate with one another, and hence show good internal consistency (Fu et al., 2023; Revelle & Condon, 2019). Even if items present different contexts or settings, or even have different levels of difficulty, internal consistency is expected if the items capture the same underlying construct (e.g. Devine & Hughes, 2016). Internal consistency is often estimated and indicated by standardised reliability coefficients, such as Cronbach's alpha and omega, that can be compared across different studies (Revelle & Condon, 2019). Another type of reliability is test-retest reliability. To the extent that mindreading is a trait-like ability (e.g., Devine, 2021), mindreading performance should be stable over short periods of time without much fluctuation in rank order and should therefore demonstrate test-retest reliability (Rust et al., 2020). Finally, when task scores are coded from open-ended responses, inter-rater reliability should be examined to ensure the scoring schemes are interpreted and applied in the same way across coders (e.g., Devine et al., 2023).

A test is considered valid if it measures the construct it is intended to capture. Validity is a matter of degree and is informed by theoretical predictions about how a given construct should behave (Nunally, 1978). Criterion-related validity concerns how well the measure predicts criterion variables, which are relevant but operationally distinct from the measure itself (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 2014), making use of the assumption that test performance has practical and theoretical implications. Hence, individual differences in test performance should be associated with behaviour, traits, psychological processes, or performance in other constructs of interest. For example, mindreading is assumed to be a keystone social cognitive ability and so should be related to social outcomes (e.g., Banerjee et al., 2011; Canty et al., 2017; Imuta et al., 2016; Devine et al., 2023). Convergent validity makes use of available measures that are viewed as measuring similar or identical constructs to the target construct under consideration. For mindreading, this might involve examining associations between measures of mindreading that use different stimuli or response formats, and other measures of social cognition. Discriminant validity is supported if the measure captures what is intended to assess, but not other constructs (Rönkkö & Cho, 2022). It is important to examine the discriminant validity of a measure as it establishes what is captured by ruling out what it does not capture.

2.1.2 The current study

Related reviews have focused on early childhood (Beaudoin et al., 2020; Fu et al., 2023; Ziatabar Admadi et al., 2015), middle childhood and adolescence, or were limited to literature that assessed alexithymia alongside mindreading (Pisani et al., 2021), or did not examine the psychometric properties of measures for adults (Derksen et al., 2018; Osterhaus & Bosacki, 2022). The current study provides the first systematic review and synthesis of measures of mindreading that have been adopted to investigate individual differences in neurotypical adults and assesses the appropriateness of measures for use in research on individual differences in mindreading performance in adults. We first summarise existing measures that have been adopted to test individual differences in mindreading in neurotypical adults. We focus on the age range of 18 to 65 because mindreading processes in older adults beyond 65 can be different from that of younger adults due to ageing (e.g., Henry et al., 2013). We analyse the evidence for the reliability and validity of each measure and examine interrelations among these measures. Finally, we discuss the differences between these measures and measures that are used to assess mindreading in children.

2.2 Method

2.2.1 Search method and selection criteria

A systematic search of relevant empirical papers published between the year 1978 (the year in which Premack and Woodruff first coined the term “theory of mind”) and January 2022 was conducted by accessing the following databases: Scopus, PsycINFO, and Web of Science on 18th January, 2022. The search terms used for searching in Scopus and Web of Science were: (“theory of mind” OR mentali?ing OR “mind reading” OR “mind perception” OR “cognitive empathy” OR “empathic accuracy” OR “mental state attribution” OR “folk psycholog*” OR “perspective taking” OR “false belief*” OR “advanced theory of mind” OR {belief-desire}) AND (adult* OR “beyond childhood” OR “lifespan” OR adolescen*). We conducted the search on PsycINFO using a combination of subject headings and search terms. We searched for entries under the subject headings “theory of mind”, “false beliefs”, or “mentalization”, in addition to those including the search terms (cognitive empathy or empathic accuracy or mind perception). The full search strategy and search timeline can be found in our preregistration on the Open Science Framework (OSF). Our search resulted in 14474 initial results published in English and other languages. After removing duplicates, 9434 papers were retained, out of which 8872 were excluded after a screening of abstracts, due to using only self-report measures, irrelevance (e.g. the search term “false belief*” generated papers referring to fallacious beliefs about the world), a focus on neural activity, absence of neurotypical adult group, or lack of availability in English. Full text of the remaining 562 papers were accessed and checked for eligibility. The final number of reports included in the review was 248, comprising of 273 studies. It was noted that some of the studies adopted more than one measure to be included in the review. The screening process is summarised in the flowchart (Fig. 2.1) following the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) statement (Page et al., 2021), and the review was preregistered on OSF prior to data analysis.

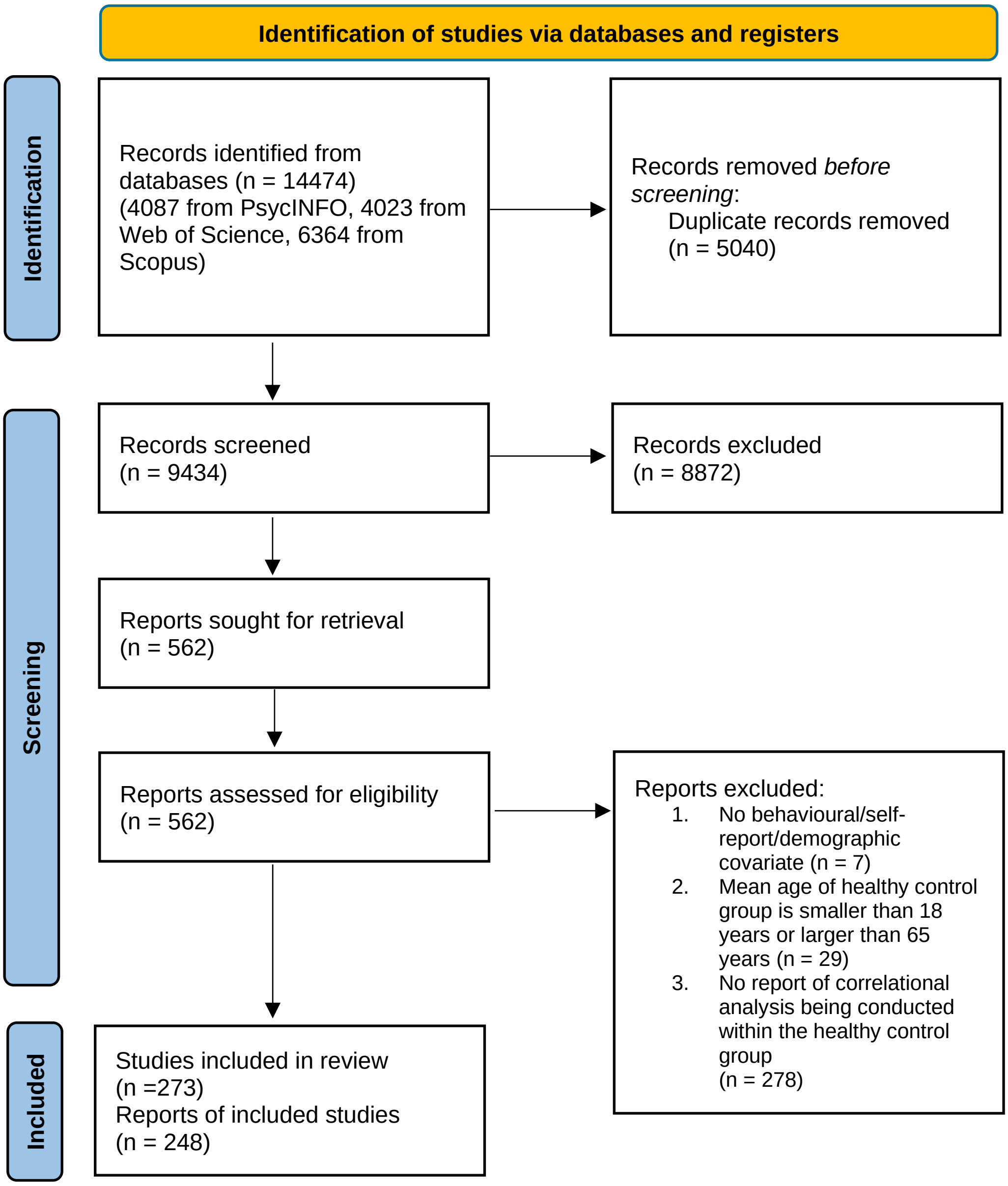

Identification of studies via databases and registers
Identification
Records identified from databases (n = 14474) (4087 from PsycINFO, 4023 from Web of Science, 6364 from Scopus)
Records removed *before screening*: Duplicate records removed (n = 5040)
Screening
Records screened (n = 9434)
Records excluded (n = 8872)
Reports sought for retrieval (n = 562)
Reports assessed for eligibility (n = 562)
Reports excluded:
1. No behavioural/self-report/demographic covariate (n = 7)
2. Mean age of healthy control group is smaller than 18 years or larger than 65 years (n = 29)
3. No report of correlational analysis being conducted within the healthy control group (n = 278)
Included
Studies included in review (n =273)
Reports of included studies (n = 248)

Figure 2.1. Flow diagram of study inclusion based on PRISMA (Page et al., 2021).

The current review focuses on behavioural measures of individual differences in mindreading in neurotypical adults. Hence, the following inclusion/exclusion criteria were adopted. We included empirical papers that included at least one group of adult participants who did not report any psychiatric or neurophysiological condition, and reported at least one correlation between mindreading performance and a behavioural, self-report, or demographic variable, or else focused on the psychometric properties of the mindreading measure(s).

Papers that only compared between groups without correlating participants' mindreading performance with any other variables were excluded to limit the number of reports included to a manageable number, and due to the reason that these papers did not contribute additional information about the test-retest reliability, convergent validity, and criterion-related validity of the reported measures without running any correlational analyses. Excluding these papers minimised the risk of overshadowing the smaller proportion of reports that showed evidence relevant to test-retest reliability, convergent validity, and criterion-related validity of the measures. Furthermore, we only included papers that included the keywords "theory of mind", "mindreading", "mentalis/zing", or "attribution of mental states" in the current review. Papers that referred to "mentalis/zation" as mind-mindedness or mind-perception were excluded, as these terms refer to the awareness or perception that other human or non-human objects have a mind without necessarily probing into the ability to infer and make use of the information about what is held in the mind of someone. We included studies that measured mindreading behaviourally and excluded studies that manipulated mindreading between different conditions, or measured mindreading in terms of neural activity. Only studies with at least one group of neurotypical participants whose mean age was between 18 and 65 years were included. Studies that examined visual perspective taking were included only if an agent with a perspective different to the participants was presented such that participants had to take the perspective of the agent, to rule out paradigms that only required mental rotation into an alternative spatial position. Studies using only self-perceived measures of mindreading were also excluded as meta-analysis results showed minimal correlation between self-reports and behavioural measures of cognitive empathy, a construct commonly defined as a component of theory of mind (Murphy & Lilienfeld, 2019). Reliability of the list of criteria was checked by a second coder screening a subset of 50 papers. The agreement between the two coders was

90%, suggesting good reliability. Discrepancy between screeners were resolved by discussion until mutual agreement was achieved.

2.2.2 Data extraction

Included papers were imported into EndNote X9 for further analysis. The major details of measures extracted included task name, the cited source of the task (task reference), stimulus type, response type, as well as scoring method and number of raters for measures using an open-ended response format. When more than one published task was merged and scored together as one larger task without distinguishing the individual components, the combined task was considered a new task. Results of the measures were also extracted, including the maximum score possible, observed range of scores, mean score, and standard deviation of scores. Psychometric properties extracted included reliability indices and any evidence of validity, and were limited to the original psychometric properties calculated from the data collected for each study. Any modifications to the measures specified by authors were also recorded. The extraction of results and psychometric properties was limited to the subset of neurotypical adult participants.

2.2.3 Coding

Information about each measure is summarised in Table 2.1. We scored the following attributes if the criteria were met in the target paper, or if they were met in the original paper from which the mindreading task was derived. The task name of each measure was unified after checking the test procedures and task references of each record. Stimulus type included stories, videos (i.e., featuring real people), photos (i.e., featuring at least a part of the faces of real people with or without context), single cartoons (i.e., single cartoon presented to prompt interpretation by participants), cartoon sequencing, animations, text in sentences, and others (e.g., interactive games). Response type included forced-choice, open-ended, sequencing, and others (e.g., pointing along a continuum). Scoring method of open-ended measures included binary scale, k-point scale (k varies from three to seven), count or proportion of certain types of response, or was not specified. The original aim of the measure first separated measures that involved testing neurotypical adults in the source paper from those that did not. For those that did, the aims were categorized into five types: population comparison (i.e., between neurotypical adults and other age groups of clinical groups), individual differences, neural underpinnings including lesion studies, experimental condition comparison, and others (e.g., norm setting). It was possible to have multiple codes for stimulus type, response type, scoring method, and original aim of measure, as the same measure may have been adapted in different ways in different studies. The tasks were coded as aiming to measure individual differences if this was explicitly stated, or if the source paper examined correlations between the task score and other behavioural or demographic variables.

Correlates were categorised into eight major types: (1) traits, with four subtypes, specifically clinical traits (e.g., autistic quotient, psychosis proneness), social traits (e.g., empathic quotient, empathic concern), personality traits (e.g., Big Five), and other traits (e.g., gender identity scale ratings); (2) social cognition measures (e.g., social intelligence, emotion recognition); (3) cognitive abilities (e.g., general intelligence, executive functions); (4) social functioning (e.g. social appropriateness, negotiation ability); (5) social outcomes (e.g. interpersonal relationship quality, intimate network size); (6) demographics; (7) miscellaneous (e.g. fatigue, fiction exposure); and (8) other mindreading measures.

The mean percent of maximum possible (POMP) score for each measure was calculated by taking the average of the mean scores in all the studies adopting the measure. In cases where it was impossible to calculate the POMP score (i.e., the mean score was presented as a raw score without reporting the maximum score possible), the entry was omitted as different studies could have adopted different scoring methods and have different maximum scores possible even when using the same measure. Where number of errors were reported and the total number of trials were reported, the mean score of the study was calculated by reversing the average proportion of error to proportion of correct responses. However, we did not calculate the POMP scores for subscales of different types of errors (e.g., undermentalising and overmentalising errors in MASC), as they reflected the type of error committed by participants rather than participants' performance.

To provide an accessible summary, reliability and validity information was coded with a three-colour system, as presented in Figure 2.2 (reliability) and Figure 2.3 (validity). Green is the most satisfactory, followed by yellow, and red indicates caution. The information was coded on a study level, as shown in Figure 2.2 and Figure 2.3, and explained below. Table 2.2 and Table 2.3 show the number of studies in which the reliability or validity of each measure was coded green, yellow, and red. The full set of extracted data are available from the link at the end of this section.

For reliability, internal consistency of a measure was coded green if the Cronbach's alpha, Guttman's lambda, or omega reported in a study was .7 or above (Cortina, 1993) or intra-class correlation (ICC) was .75 or above (Fleiss, 1986); it was coded yellow if alpha/lambda/omega indices were between .6 and .7, ICC was between .5 and .75, or split-half reliability was between .5 and .75. If different indices in the same study conflicted in colour coding, the coding was decided upon the value of the alpha/lambda/omega index. Test-retest reliability was coded green if the correlation coefficient between two time points administering the same test within eight weeks was .70 or above or intra-class correlation (ICC) was.75 or above, yellow if the correlation was between .4 and .70 (.75 for ICC), and red if the correlation was below .4 (Cicchetti, 1994; Fleiss, 1986). Inter-rater reliability was coded green if the Cohen's Kappa or intra-class correlation was .75 or above (Mordal et al., 2010); average indices between .4 and .75 were coded yellow and those below .4 were coded red. An observed factor structure being consistent with the one hypothesised was taken as evidence supporting the factor structure of the measure. Most of the time, the measures proposed to capture a unitary mindreading component, and the factor structure was supported if the results showed a good fit to a one-factor model. In other measures that included a control scale or proposed several subscales, a good fit to a two-factor model that distinguished the mindreading subscale and the control subscale, or the proposed subscales, were treated as evidence for the proposed factor structures.

Validity was colour-coded based on whether the studies reported evidence for or against different kinds of validity. Green was coded when there was only supporting evidence within a single study; yellow referred to mixed evidence within a single study (i.e. having both evidence that supports and opposes validity in the same study, such as reporting one correlation larger than the effect size threshold we will later specify, and another correlation smaller than the threshold), and red was coded when there was only evidence against validity in the specific way, within a single study. We coded for four types of validity evidence, conceptually similar to convergent validity, criterion-related validity, known-group validity and discriminant validity.

We coded for “broad” convergent validity and “narrow” convergent validity. Reports of performance on the measure correlating with other social cognition or social ability measures, not limited to mindreading, were taken as evidence of broad convergent validity. Positive evidence was characterised by a Pearson’s or Spearman’s correlation coefficient of .19 (taking the absolute value) or higher, which is the median effect size in individual differences studies (Gignac & Zoderai, 2016). By adopting this criterion, which is less stringent than Cohen’s convention of .30 for a medium effect size (Cohen, 1992), we expect to err on the side of an optimistic picture of convergent validity displayed by the identified tasks. Correlations of task performance and general social abilities or relevant clinical traits, specifically autistic quotient (AQ) or alexithymia trait scores, were also included as evidence regarding broad convergent validity, for the questionnaires include components that tapped on social cognitive abilities. The same .19 threshold explained above was applied in such cases. In most cases evidence in favour of convergent validity came from positive correlations, but it was also possible for negative correlations to provide positive evidence (e.g., when one of the correlated measures examined response time, or when participants' mindreading performance was correlated with clinical traits associated with social difficulties). For narrow convergent validity, we investigated interrelations among the mindreading tasks identified in this review for relevant evidence. Two tasks were taken as correlated in a study if there was at least one correlation that exceeded the .19 threshold between any subscales of the two tasks. Any lower correlations reported in studies were considered evidence against interrelation between two tasks.

Criterion-related validity was supported by evidence suggesting a correlation between performance on the measure and social functioning or social outcomes (e.g., interpersonal relationship quality, community functioning, social functioning scale performance). Known-group validity was supported by reports of differences in performance on the measure between the neurotypical adult control group and clinical groups showing social deficits, specifically autism spectrum disorder (ASD) and schizophrenia, or between participants grouped by high versus low autistic or schizophrenic traits, or either children or older adults. Discriminant validity was supported by results showing that (1) the measure contributed to unique variance in criterion variables including social functioning and social outcomes after controlling for at least one of three confounds: verbal ability, general intelligence, executive functions; (2) only the subscale(s) relevant to mindreading but not the control subscale(s) correlated with the criterion variables; (3) known-group differences in task performance remained significant after controlling for at least one of the three confound variables; (4) known-group differences in the mindreading-relevant and control subscales were dissociated;

or (5) known-group differences in mindreading-relevant subscale(s) remained significant after controlling for the scores on control subscale(s).

The full set of extracted data and the spreadsheets for coding the data are publicly available on OSF (https://osf.io/23ynq/?view_only=7f34abba115b40da99c14b1e08d97f67).

2.2.4 Sample characteristics

Approximately 47640 neurotypical participants aged between 18 and 65 were included in the 273 studies. The smallest study had 10 participants and the largest study included 2242 participants. The average sample size was 173 (around 62% female with all samples aggregated, excluding studies that did not report gender).

Twenty studies did not report the mean age of participants. The mean age of participants in the remaining 253 studies varied from 18.12 years to 59.27 years, and the average of mean age reported in studies was 30.04 years.

2.3 Results

We begin by describing the key features of the stimuli and measurement formats of the tasks identified. Next, we evaluate the psychometric properties of the tasks, with particular focus on the eight tasks for which we have the most data to inform evaluation. We also evaluate the interrelations among the measures identified.

2.3.1 Description of identified measures (Table 2.1).

We identified 75 measures that have been adopted to assess individual differences in mindreading in neurotypical adults, including one unpublished measure with no further information, listed in Table 2.1. The mean age of participants is also summarised in Table 2.1. Forty-three (57%) measures were designed for detecting differences between groups in adults (e.g., adults with a known diagnosis vs. those without a diagnosis) rather than individual differences; 26 (35%) were designed to detect individual differences in adults. The mean age of participants ranged from 19.50 to 49.60 years, with an average of 30.14 years.

Forms of stimuli. Out of the 75 identified measures, many of the measures involved narratives or stories (52; 69%) presented as text or speech (27; 36%), videos (15; 20%), cartoon sequences (11; 15%), or animations (4; 5%). Two animation tasks featured geometric shapes rather than human agents. The forms of stimuli adopted by the remaining measures are listed in Table 2.1. The types of stimuli presented in five (7%) tasks were inconsistent across studies (e.g., for the Hinting task some studies presented narratives while some presented videos). How the participants were required to respond to the stimuli, and how their responses were measured, are discussed next.

Form of measurement. There was considerable variety in measurement methods, not only between tasks, but also when the same notional task was used in different studies. This limits the confidence with which conclusions about reliability and validity from a study using one task variant can be expected to generalise to studies using another task variant.

Response format. Most of the measures involved forced-choice responses and/or open-ended questions. Among the 75 measures, 45 (60%) involved a forced-choice between two and five alternatives, 31 (41%) required open-ended verbal responses, four (5%) involved subjective ratings (e.g., rating the likelihood of possible explanations to an agent's behaviour, or the likelihood of an agent having different emotional responses in a described social scenario), three (4%) involved picture sequencing, one (1%) involved pointing to a location within a continuous space, and one (1%) required moving a designated object as directed. Four measures (5%) involved at least two components (e.g., including both sequencing and open-ended questions). The response formats were inconsistent across studies for seven measures (9%), and one additional measure (1%) had a different number of forced-choice options in different studies.

Scoring method. As forced-choice and open-ended responses were the two most popular response formats, this subsection describes how the items were scored across different studies using the same measure. The analysis revealed considerable diversity between different methods, and between different studies using the same method.

As for forced-choice measures, dichotomous scoring that differentiated correct from incorrect answers for items was used in 39 (87%) of the forced-choice measures, while eight measures (18%) involved scoring on a k-point scale (k varies from three to seven) that rated participants' item responses according to the extent they matched with developed scoring schemes. One measure (2%) weighted scores by expert ratings of an agent's possible mental states that can arise from a described social scenario, which was collected a priori. Among the 45 measures that involved a forced-choice response format in at least one study, four (9%) have been scored using more than one of the above methods across studies.

For the 31 measures that were used with an open-ended response format in at least one study, twenty (63%) measures scored open-ended items on a k-point scale (k varies from three to seven), according to how much the participant's response matched a developed coding scheme. Fourteen measures (45%) adopted dichotomous scoring (correct or incorrect). Four measures (13%) scored participants' performance by counting or calculating the proportion of mental state references in their responses. Scoring procedures for three measures (10%) using open-ended items were not reported. Six measures (19%) were scored on more than one dimension, and 10 (32%) were scored using inconsistent methods in different studies.

Most open-ended measures were scored either according to correctness of responses, or/and evidence of a propensity to mentalise. Within the 25 (81%) open-ended measures that scored responses based on correctness, 18 (72%) scored responses on a non-binary scale and thus allowed for partial scoring. One or more of the following criteria were used to judge the score to be awarded: order of inference, extent of explicit mental state description, contextual relevance, the number of times the experimenter gave a prompt, and explanatory power.

Seven (23%) open-ended measures captured participants' propensity to mentalise on a binary scale indicating whether the response involved mental state attribution (2 measures; 29%), a 3-, 5-, 6-, or 7-point scale reflecting the degree of deliberateness of mental state attribution (2 measures; 29%), or the occurrence of mental state references in the participants' responses in terms of count or proportion (4 measures; 57%).

Within the two measures (6%) that did not score responses on correctness or propensity, one measure scored responses on their coherence, clearness and abundance of contextualised examples; one measure did not specify the scoring criteria.

2.3.2 Ceiling effects and psychometric properties of measures

We first summarise the overall availability of relevant evidence from all 75 measures (see OSF for full data). Many tasks have only been used in a small number of studies, and many studies did not include evidence relevant to ceiling effects or psychometric properties. We therefore proceed to a more detailed evaluation on the eight tasks that have been used to study individual differences in neurotypical adults in 10 studies or more. As will become clear, even for these measures there is only limited evidence about reliability and validity, and we judged it even less likely that it would be possible to draw conclusions on the psychometric properties of measures where even less information was available.

Sensitivity to individual differences in performance. Where relevant data were available there was considerable evidence of ceiling effects. We report mean Percentage of Maximum Possible (POMP) scores and POMP score ranges to identify ceiling effects in Table 2.4. Table 2.4 shows the mean POMP scores and range of POMP scores for all measures. A task is sensitive to individual differences in a population within a particular age range when the POMP score is within the range of 20% to 80% (e.g., Petersen et al., 2016). We used 85% as the cut-off for indicating a ceiling effect to allow for more leniency. Measures that show a ceiling effect for at least one of the subscales are highlighted in red, including 29 measures (49% of measures that have available POMP score information) based on mean POMP score, and 13 measures (50% of measures that have available information on POMP score range) based on POMP score range. Nine measures (12%) did not have information about their mean POMP scores available because mean scores or maximum possible scores were not reported, and POMP scores were not applicable for seven measures (9%) due to their response formats (e.g., measures involving only reaction time, measures that calculated scores by taking the differences between ratings, measures that counted the number of mental state utterances). Range of POMP scores were not available for 51 (68%) measures, mostly because the measures were only used in one study.

Summary of reliability and validity reports. Among all 75 measures, 30 (40%) did not have information about reliability and 20 (27%) did not have information about validity (beyond face validity). Evidence of internal consistency was available from at least one study for 34 measures (45%). Evidence regarding factor structure was available for 16 (21%) measures. Only 6 (8%) measures had evidence for test-retest reliability. Evidence of inter-rater reliability was available for 16 out of 31 (52%) measures that were conducted in open-ended format in at least one study. Evidence regarding broad convergent validity was reported at least once for 49 (65%) measures, while there was evidence of known-group validity for 29 (39%) measures. Additionally, evidence of discriminant validity was available for 17 (23%) measures, and evidence regarding criterion-related validity was available for 9 (12%) measures.

Narrow convergent validity: Interrelations among measures. We examined the interrelations among mindreading measures identified. Twenty-nine (39%) measures had no data bearing on their correlations with other measures. Two (4%) of 46 measures correlated with other mindreading measures were not included in the analysis of this section as the correlations were not conducted specifically in the neurotypical adult group. Table 2.5 shows the interrelations among 44 measures (59% of 75 measures) for which there was relevant evidence, 43 of which had at least one correlation coefficient reported. When multiple correlations were conducted between different subscales or versions of the same task within the same study, we made our evaluation of positive evidence on the basis of the maximum correlation coefficient reported (taking the absolute value). This approach allowed us simplify and present the most optimistic picture of the overall correlation patterns among measures.

In total, there were 98 correlations reported, 93 (95%) of which also specified the value of the correlation coefficient. We applied a threshold of .19 for Pearson's correlation or Spearman's correlation. Out of the 93 correlations with reported coefficient values, 63 (68%) exceeded the cut-off. Among the 43 measures, 10 measures (23%) showed correlations with other measures that had an effect size smaller than the threshold.

The "top 8" measures. We investigated the properties of the eight tasks that were used most widely in published research. These eight measures comprised the RMET (Baron-Cohen et al., 2001; 149 studies), Strange Stories Task (Happé, 1994; 33 studies), Faux Pas Recognition Task (FPRT; Baron-Cohen et al., 1999; 28 studies), Hinting Task (Corcoran et al., 1995; 25 studies), ToM Picture Stories Task (Brüne, 2003; 12 studies), Movie for the Assessment of Social Cognition Task (MASC; Dziobek et al., 2006; 11 studies), Imposing Memory Test (Kinderman et al., 1998; 11 studies), and Animations Task (Abell et al., 2000; 10 studies). Even among these tasks, reporting of information related to reliability and validity was infrequent. The highest rate was 16 out of 33 studies employing the Strange Stories task reporting inter-rater reliability, and rates were generally much lower (Table 2.2). Consequently, the data available to evaluate reliability and validity is limited, and comes disproportionately from one task, the RMET. This is important to keep in mind when evaluating the summary diagrams in Figures 2 and 3.

For the “top 8” measures, ceiling effects were shown in participants’ average performance on three tasks: Strange Stories Task, FPRT, and both components of ToM Picture Stories Task as well as its total score. The minimum POMP score reported for the total score on the ToM Picture Stories Task (89.42% among six studies) also exceeded the 85% cut-off.

Table 2.2 and Figure 2.2 list the eight measures and the availability of information on their reliability, in alphabetical order. It should be noted that information about inter-rater reliability is only available for measures that have been used with an open-ended format in at least one study, including Animations Task, FPRT, Hinting Task, and Strange Stories Task. It was noted that inter-rater reliabilities of RMET were reported in two studies in which the tasks were presented in a forced-choice format, but we do not include this information in the current summary because reports of inter-rater reliability of forced-choice measures are not informative. There was evidence regarding internal consistency for all eight measures. Table 2.6 shows the average Cronbach’s alpha of the top eight measures, and the Hinting task is the only task that had an average Cronbach’s alpha falling below 0.6. Five tasks had evidence for factor structure, whereas evidence regarding test-retest reliability was only available for the Hinting Task and the RMET, and this evidence was mixed.

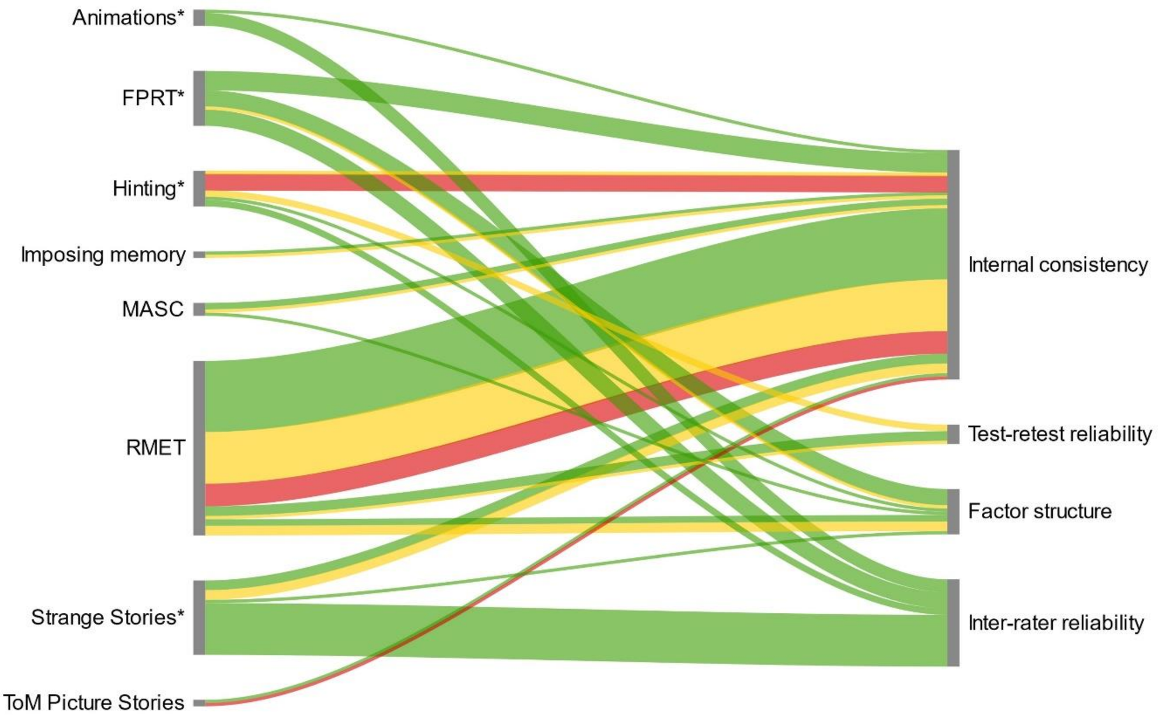


Figure 2.2. Available evidence regarding reliability of the top 8 measures.

The diagram depicts the availability of evidence for or against reliability of the top eight popular measures, including Animations Task (Animations*), Faux Pas Recognition Task (FPRT*), Hinting Task (Hinting*), Imposing Memory Test (Imposing Memory), Movie for the Assessment of Social Cognition Task (MASC), Reading the Mind in the Eyes Test (RMET), Strange Stories Task (Strange Stories*), and ToM Picture Stories Task (ToM Picture Stories), in alphabetical order. The tasks that were presented in an open-ended response format in at least one study were indicated with "*". The colour coding follows the same principle as for Table 2.2, with green indicating the most satisfactory evidence according to standard criteria, yellow intermediate, and red the least satisfactory. Curve width is weighted by number of studies showing relevant evidence for or against reliability. Curves extended from the same measure should have equal width if the same number of studies indicate evidence for or against the specific type regarding reliability of the same measure.

Table 2.3 and Figure 2.3 list the eight measures that have been adopted in 10 studies or more and the availability of information regarding their validity. All eight measures had evidence regarding broad convergent validity. Positive evidence was most frequent, but evidence was mixed for 6 of 8 tasks and only negative for one (Animations Task). We extended our analysis of narrow convergent validity to the calculation of interrelations among these eight measures by applying correction for attenuation, to reduce the potential underestimation of interrelationships stemming from the measures' less-than-perfect internal consistency. This correction was possible for the top eight measures as reported values of Cronbach's alpha were available and could be averaged for each measure (see Table 2.6). Twenty-seven (93%) of the 29 correlations between the top eight measures had correlation coefficients reported, 18 (62%) and 21 (78%) of which exceeded the threshold of .19 before and after the correction, respectively. Table 2.7 lists the correlation coefficients among the top eight measures, and the number of studies that reported at least one relevant correlation that exceeded the .19 threshold, before and after correction of attenuation.

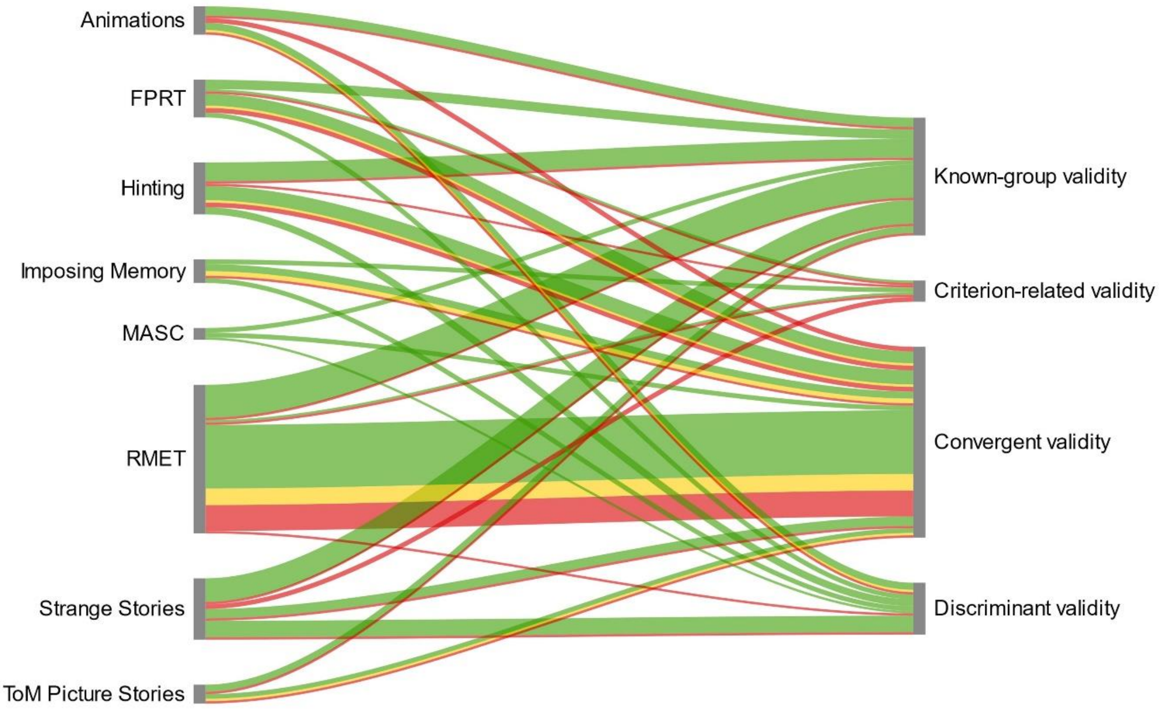


Figure 2.3. Available evidence regarding validity of the top 8 measures.

The diagram depicts the availability of evidence for or against validity (beyond face validity) of the top eight popular measures. The colour coding follows the same principle as for Table 2.3. Curve width is weighted by number of studies showing relevant evidence for or against validity. Curves extended from the same measure should be equal in width if the same number of studies indicate evidence for or against the specific type regarding validity of the same measure. Convergent validity in this diagram refers to broad convergent validity.

Seven out of eight tasks have some evidence regarding discriminant validity. Most of this evidence was positive, though at low frequencies. The number of studies providing evidence relevant to criterion-related validity of these measures was especially limited, with only 9 studies, and only 4 of these providing positive evidence. Notably there was no evidence regarding criterion-related validity for the Animations task, the MASC, or the ToM Picture Stories Task.

2.4 Discussion

The current systematic review considered measures that have been used to examine individual differences in mindreading in neurotypical adults, specifically identifying the basic characteristics of the tasks, and examining ceiling effects, reliability and validity of the measures, employing a systematic strategy. We evaluated the measures with reference to established psychometric criteria, and observed that no current measure provided strong, consistent evidence of robust psychometric properties. We summarise these findings below, compare the identified measures with mindreading measures for young children, make recommendations for the conduct and reporting of future research using existing measures, and identify the need to further examine psychometric properties of existing research and develop new measures that are more likely to show good psychometric properties.

2.4.1 Description of identified measures and standardisation of administration

Only one-third of the identified measures were specifically designed to study individual differences. Of course, tasks designed for other purposes may nonetheless succeed in measuring individual differences, but this cannot be taken for granted, and the high proportion of tasks designed for other purposes may explain evidence of poor psychometric properties. Most of the tasks employed a forced-choice response format. Open-ended responses were also common, but inter-rater reliability was not consistently reported. Moreover, while most tasks focused on scoring the correctness of responses, a few assessed participants' propensity to make mental state attributions irrespective of correctness. This observation suggests a lack of consensus about how to operationalise individual differences in mindreading. It is currently unclear whether there might truly be multiple sources of individual differences in mindreading, or just incidental variation in methods.

The tasks varied in terms of stimuli and measurement formats, and tasks that were notionally the same were often implemented with different stimuli or scoring criteria between studies. While each individual study can nonetheless be evaluated on its own merits, these inconsistencies complicate the comparison of participants' performance between studies or measures. It also means that the psychometric properties of an adapted task cannot be inferred from other studies using the original version of the task (nor vice versa). For example, drawing from research on young children, research by Hughes et al. (2000) showed that the good test-retest reliability of standard false beliefs tasks was masked by the nonstandard approach of administration by Mayes et al. (1996). Similar effects are plausible in testing neurotypical adults as well.

2.4.2 Inspection of ceiling effects and psychometric properties

Psychometric theory provides criteria for evaluating reliability and validity, which bear on the ability of a test to measure a psychological construct (e.g., Rust, Kosinski, & Stillwell, 2021). For research on individual differences, tests must be sensitive to variation without evidence of ceiling and floor effects. A test must also show internal reliability (whereby a participant who performs well on one item tends also to perform well on other items measuring the same construct), without which it is unclear that test scores are informative about any underlying construct. It is also highly desirable that a participant who performs well on one occasion is also likely to perform well if tested later (i.e., the test shows test-retest reliability), because this indicates stability in how well the test captures the underlying construct over repeated measures. It is, of course, possible to have a highly reliable test that shows low validity because it fails to test the intended psychological construct. To evaluate validity, it is common to consider whether a test correlates with other tests of the same construct, whether it correlates with tests of other abilities, behaviours, or outcomes relevant to the construct, and whether the test is sensitive to differences between groups that differ in those abilities, behaviours or outcomes. It is also important to distinguish what a test measures from other distinct but relevant constructs. We will summarise our findings against each of these criteria.

Ceiling effects. Around half of the tasks showed a ceiling effect for at least one subscale (as evidenced through percentage of maximum possible scores), indicating that many tasks did not generate enough variance to study individual differences in neurotypical adults effectively. Adopting such measures can lead to erroneous conclusions that there are no individual differences in mindreading in adults due to the insensitivity of the measure rather than the absence of meaningful differences in the underlying ability (e.g., Anastasi, 1948). When there is little variance within the sample, the limited spread of unique values makes it harder to detect relationships between participants' performance on the measure and other variables. While techniques for correcting range restrictions can help mitigate the underestimation of correlations with other variables, other issues, such as skewed distributions of scores, still exist, which might provide a distorted picture of the relationship between task performance and other variables of interest. Thus, mindreading measures with marked ceiling effects in a target population (i.e., where the average score is > 80% of maximum possible score) are unsuitable for measuring individual differences (e.g., Petersen et al., 2016).

Reliability and validity. Information on reliability and validity was often not reported, even among the eight mindreading tasks that were adopted most frequently. Available data showed that seven out of the top eight tasks had at least acceptable internal consistency (the Hinting task was the exception). This provided support for the claim that the items in a given task reliably captured a single construct (i.e., mindreading). A point to note is that good internal consistency of a task does not preclude that items vary in difficulty, or that success requires participants to adapt their reasoning to the context of individual items, as items are expected to be correlated with one another if they capture the same underlying construct. Apart from internal consistency, there was also mixed but acceptable evidence supporting inter-rater reliability and factor structure. However, very few tasks had information on test-retest reliability. If we assume that mindreading is a stable trait, examining test-retest reliability is important to show that the task is tapping on the construct rather than a state that varies over time (Matheson, 2019).

As for validity, known-group validity and discriminant validity were generally satisfactory for the top eight tasks, with the exception that there was no reported evidence for known-group validity and discriminant validity for the Imposing Memory test and the ToM Picture Stories task, respectively. There was more abundant evidence regarding convergent validity for the top eight tasks, but the evidence was mixed for six tasks (except for the MASC and the Animations task). There was only evidence that support good convergent validity of the MASC, but there was no evidence for good convergent validity of the Animations task. There was especially limited information about criterion-related validity of the measures. This is a striking limitation of current literature, which means that, whether or not current tasks are measuring mindreading reliably, there is little evidence (positive or negative) that they are measuring something that "matters" for social behaviour, mental health, or wellbeing.

Unsurprisingly, there was more information available regarding psychometric properties of tasks that are more frequently used. It is imperative to establish psychometric properties first, such that researchers have enough information to make informed decisions. For example, the RMET, being the most frequently used measure, had the most evidence for evaluating its psychometric properties. However, results showed that it did not exhibit the best reliability or validity. This can be because the small number of studies that adopted other tasks exaggerated the appearance of consistent evidence. Nevertheless, some tasks may demonstrate strong psychometric properties, yet lack sufficient supporting evidence due to their infrequent use. What is needed is consistent reporting of psychometric properties to generate a larger evidence base. It is suggested that researchers refer to existing guidelines on reporting psychometric properties of measures, for example, The Standards for Educational and Psychological Testing (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 2014).

Interrelations among measures. We examined the interrelations among the identified measures to investigate convergent validity. We found inconsistent evidence of intercorrelation, and some measures were not correlated with any other measures. This may reflect ceiling effects and unsatisfactory reliability of some measures, but also the possibility that mindreading may be multi-dimensional rather than uni-dimensional. In the case of problematic ceiling effects, applying correction for attenuation to the interrelations among the top eight tasks did not change the overall picture, as only three correlations that fell below the .19 threshold before correction exceeded the threshold after correction. This observation implies that the lack of interrelations among tasks cannot be fully attributed to reliability issues. Another possible reason for the mixed interrelations is range restriction due to limited variance in task performance, as explained above in the discussion of ceiling effects, which might have masked genuine underlying associations among the tasks (Mendoza & Mumford, 1987); range restriction can also occur when some samples are highly homogenous, for instance, when assessing only university undergraduates within a single sample. This could also be a reason why we found mixed evidence for broad convergent and criterion-related validity of tasks that exhibited ceiling effects. The lack of interrelations among certain tasks might also be attributed to attenuation of correlations due to distinct task demands for different tasks. A latent variable approach is one way of addressing this problem of task impurity: if a common latent factor emerges this provides evidence that the tasks capture a common construct despite having different incidental requirements.

Moreover, the inconsistency of interrelations among tasks might reflect multidimensionality of mindreading. Mindreading is a loosely defined construct with diverse operationalisations (Apperly, 2010; Happé et al., 2017; Schaafsma et al., 2015; Warnell & Redcay, 2019). While all tasks reviewed had face validity as mindreading tasks, researchers need to look beyond face validity, because superficial resemblance to the construct of interest does not guarantee accurate and specific assessment. For example, despite the face validity of the RMET there is evidence that this task measures emotion perception rather than theory of mind (Oakley et al., 2016). This issue particularly warrants concern when considering that different tasks require participants to engage in different activities, including but not limited to making mental state inferences about characters from vignettes, photos and videos, interpreting non-literal speech, and recognising social transgressions. Face validity does not elucidate whether a task in fact captures a common underlying construct. While studies using latent variable analysis have identified a single underlying latent construct of mindreading in early childhood, middle childhood and adolescence (e.g., Devine et al., 2023; Hughes, Devine, & Wang, 2018), similar work with adults has yet to be undertaken.

Another possible reason for inconsistent associations among tasks is that some tasks may not index mindreading ability. It is difficult to establish if a task captures mindreading or not when researchers have not mapped out the taxonomy of abilities that make up the construct of mindreading. Some literature has suggested useful theoretical principles to distinguish whether a task captures mindreading, for example, the necessity to represent mental states and distinguishing one's own mental states from that of others (Quesque & Rossetti, 2020). However, tasks that fulfil such criteria might be measuring only a specific sub-ability of self-other distinction under the general latent construct of mindreading, which might include motivational as well as structural components. Therefore, it is imperative for mindreading researchers to tackle theoretical issues regarding the nature of mindreading in adults.

2.4.3 Use of measures of mindreading for children

In the current review we observed that tasks designed for testing developmental differences or individual differences in young children show ceiling effects in adults. It should not be surprising that tasks designed to test basic possession of mental state concepts – such as false belief tasks – show little variation in performance among participants who are far older than the age at which children typically pass these tasks. This is supported by our findings, which suggest that these tasks should not be used to study individual differences in adults.

A substantial number of the studies reviewed here adopted tasks originally designed to be “advanced” tests of mindreading in older children and adolescents. These tasks are sometimes also more naturalistic, bearing higher resemblance to reality where using mindreading is more complex and dynamic, compared to laboratory tasks that only focus on specific mental state concepts. Two measures designed for older children, the FPRT and unexpected outcome test, showed different results. The FPRT exhibited a ceiling effect, while the unexpected outcome test did not, although the POMP score calculated for the latter was based on just one study. Other popular tasks have been used for testing older children, such as the Strange Stories task, Animations task, and Hinting task. Some of these tasks show ceiling effects in adults, while others did not (refer to Table 2.4). It is worth noting that RMET has a child version with fewer items and simpler vocabulary, specifically designed for testing children. Tasks like RMET and Hinting task can be useful for studying how mindreading abilities develop from childhood to adulthood and have the potential to provide insight into the continuity of mindreading across lifespan. In summary, some tasks originally designed for older children show promise as measures of individual differences in adults. However, like the tasks designed for adults it is unclear what these tasks measure beyond variation in "mindreading".

2.4.4 A programme for future work

The current literature provides considerable prima facie evidence of individual differences in mindreading in adults, but much more limited evidence that these differences are psychometrically robust, surprisingly little insight into what this variation might mean, and little evidence that mindreading matters for social outcomes in neurotypical adults. New conceptual work and conceptually-motivated empirical work is necessary to clarify in what sense people vary in mindreading abilities after they pass the standard assessments of mental state concepts that have been devised for children (e.g., the concepts of desire or belief). Likewise, conceptually-motivated work is necessary to develop a taxonomy of potential mindreading components and support the selection of tasks that target such components (Apperly, 2010; Happé et al., 2017; Schaafsma et al., 2015; Warnell & Redcay, 2019). This is likely to require the development of new tasks as well as the systematic examination of existing tasks. In both cases it is essential that the field move towards consistent reporting of information for establishing reliability and validity of measurement. If tasks require component abilities, then examining convergent and discriminant validity is critical to test whether this is reflected in individual differences in performance. The most powerful way to do this is to collect data from multiple tasks in the same participants and test theoretically motivated models of the co-variance. Empirical support for sub-components of mindreading would come from meeting two conditions. First, tasks targeting each sub-component should load onto distinct latent variables (demonstrating convergence between tasks testing that sub-component, and divergence from tasks testing other sub-components); second, latent variables for sub-components should nonetheless be correlated (Devine, 2021). Meeting this second condition supplies empirical grounds for saying that the latent variables measure sub-components of a common underlying construct (i.e., mindreading). Such a pattern would be similar to findings reported in the executive function literature, which shows shared variance across latent variables that tap on different subdomains, including inhibition, shifting and updating (e.g. Friedman & Miyake, 2017; Miyake et al., 2000). Mapping out the taxonomy of sub-components will help to elucidate the nature of individual differences in adults' mindreading.

Finally, it is clearly important to establish that such variance in adults matters for relevant outcomes in real social behaviour, mental health, or wellbeing as much as it appears to matter in childhood (e.g., Hughes & Devine, 2015). The current literature provides a considerable amount of evidence of known-group validity – demonstrating that neurotypical adults perform at higher levels on a given mindreading task than a clinical group that is known to have social difficulties. This is clearly of considerable value and interest, but it does not demonstrate that variation in mindreading matters for people who do not have a clinical diagnosis. Such evidence is almost entirely lacking at present, and so testing this criterion validity for individual differences in mindreading in adults is a clear priority for future work.

2.4.5 Implications

This review can be used as a reference tool for researchers from all disciplines in psychology who want to examine individual differences in mindreading in neurotypical adults to select appropriate task(s). We also suggest a list of attributes concerning reliability and validity that researchers should report when they adopt any of the measures to facilitate future systematic review work in the field, or even meta-analyses. Moreover, the investigation on interrelations among tasks informs us of the potentially multifaceted domain structure of mindreading.

2.4.6 Limitations

One limitation is that we only included English papers for the current review, which may have excluded relevant studies published in other languages. Another limitation is that many measures reviewed lacked comprehensive report of psychometric properties, which limits the confidence of our synthesised results, as it is important to note that lack of evidence is not evidence of absence. Moreover, the current review does not delve into the contentious topic of operationalisation of mindreading. We included all measures that purported to be assessing mindreading, because our primary objective was to inspect the psychometric properties of such measures. Furthermore, we did not review task durations; measures with good psychometric properties may not be suitable for certain research contexts where time allowed for data collection is limited. Another limitation is that we did not evaluate the relevance of tasks identified to the participants. For example, based on the limited available evidence the MASC shows satisfactory psychometric properties and does not show ceiling effects. However, the video stimuli involve a dinner-date scenario between three white, apparently middle-class Germans aged around thirty to forty. For people who do not speak German it is commonly dubbed into English. While the demographic specificity may help with the realism of the scenario, it also raises the realistic possibility that participants' understanding of the scenario will vary depending upon their own demographics, that is, the task may not demonstrate measurement invariance. This serves to illustrate the general point that it cannot be assumed that the psychometric properties of a test are fixed across contexts. Instead, measurement invariance needs to be established in diverse settings (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 2014; Nunally, 1978).

2.4.7 Future directions

Our findings show that further research on psychometric properties of mindreading measures is necessary. We suggest two ways for relevant investigation in the future: the first way is to conduct further research on examining and improving current measures, and the second way is to design new measures that exhibit better psychometric properties.

Recommendations for new research with existing measures. We recommend that measures that exhibit ceiling effects in children should not be used for testing adults. Researchers should always check for ceiling effects. We suggest that more studies that focus on examining psychometric criteria of existing measures be done, and studies adopting such measures should report evidence on reliability and validity. When measures with less satisfactory reliability are adopted, we suggest the use of multiple measures with latent variable modelling to better partial out measurement errors. By using latent variable modelling, the relationships among measures can also be evaluated.

Recommendations for the development of new measures. New measures should aim to achieve good reliability and validity. It is also important to ensure that the measures are relevant and suitable for the participants of interest; age range and culture of participants should be taken into consideration.

2.4.8 Conclusion

The current review highlights a large evidence gap, whereby the great majority of studies that have examined individual differences in mindreading have not examined whether the tasks are either reliable or valid. In some cases, this is problematic, such as where ceiling effects preclude any meaningful conclusions. The picture emerging from existing evidence provides only very limited confidence in the measurement properties of existing measures, highlighting the need to gain further evidence of reliability and validity of existing measures and to consider development of new measures. Interrelations among measures were inconsistent, which could be due to measurement problems, or due to tasks measuring different aspects of mindreading. This highlights the need for empirical work to be aligned with theoretical work on the origins and structure of individual differences in mindreading in adults, which should inform both the development of new tasks, and more precise hypotheses about the relevance of mindreading for social abilities, mental health and wellbeing.

Tables

Table 2.1. List of measures identified (in descending order of occurrences in studies). The “top eight” measures discussed in most detail in the text are shaded.

Measure name	Task reference	No. of studies	Original aim	(Range of) mean age	Stimulus type	Response type	Item scoring method (* refers to scoring method used in the original reference; # refers to total score)	Scoring attribute
Reading the Mind in the Eyes Test	Baron-Cohen et al., 2001	149	Population comparison (clinical); individual differences	Range: 18.1-59.2 Mean: 29.0	Photos (eyes)	Forced-choice (3/4 options)	Binary scale	Correctness
Strange Stories Task	Happé, 1994	33	Population comparison (clinical)	Range: 18.6-47.7 Mean: 28.6	Stories	Open-ended	Binary scale/3-point scale	Correctness
Faux pas recognition test	Baron-Cohen et al., 1999	28	* Population comparison (clinical); task comparison (designed for children)	Range: 18.6-59.2 Mean: 32.8	Stories	Open-ended	Binary scale/3-point scale	Correctness
Hinting task	Corcoran et al., 1995	25	Population comparison (clinical)	Range: 20.1-51.7 Mean: 31.6	Stories/ Videos	Open-ended	Binary scale/3-point scale/4-point scale	Correctness

ToM Picture Stories task	Brüne, 2003	12	Population comparison (clinical)	Range: 20.5-46.3 Mean: 34.0	Cartoons (sequence)	Forced-choice (3 options)/Sequencing & Open-ended	7-point scale (sequencing); n/a (sequencing time); #23 max (open-ended questionnaire total score)	Correctness; n/a; correctness
Imposing memory test	Kinderman et al., 1998	11	Population comparison (group split by other variables)	Range: 20.3-53.0 Mean: 28.8	Stories/ Videos	Forced-choice (binary)	Binary scale	Correctness
MASC	Dziobek et al., 2006	11	Population comparison (clinical); individual differences	Range: 19.9-47.0 Mean: 28.6	Videos	Forced-choice (4 options)	Binary scale	Correctness/ (propensity if taking into consideration the type of error committed)
Animations task	Abell et al., 2000	10	Population comparison (clinical)	Range: 19.3-32.3 Mean: 24.9	Animations	Forced-choice (4 options)/Open-ended	Binary scale/3-point scale/*6-point scale (intentionality subscale)	Correctness/propensity
False belief task (1st-order + 2nd-order)	Perner & Wimmer, 1985	8	*Developmental differences (designed for children)	Range: 21.9-35.5 Mean: 27.1	Cartoons (sequence)/Stories	Forced-choice (binary)/Open-ended	Binary scale	Correctness
TASIT	McDonald et al., 2003	8	Population comparison (clinical)	Range: 19.7-40.7 Mean: 29.6	Videos	Forced-choice (binary/3 options)	Binary scale	Correctness
Yoni task	Shamay-Tsoory & Aharon-Peretz, 2007	6	Population comparison (clinical)	Range: 19.8-25.9 Mean: 22.7	Illustrated items	Forced-choice (4 options)	Binary scale	Correctness

Short Story Task	Dodell-Feder et al., 2013	5	Individual differences	Range: 19.4-27.8 Mean: 23.6	Stories	Open-ended	Binary scale (spontaneous subscale); 3-point scale (explicit mental subscale)	Correctness/propensity
Director task	Keysar et al., 2000	4	* Experimental condition comparison (age not mentioned)	Range: 19.1-23.0 Mean: 21.3	Interactive game	Action	Binary scale (error measure); n/a (RT measure)	Correctness; n/a
Picture sequencing task	Langdon et al., 1997	4	Population comparison (clinical)	Range: 32-47.7 Mean: 40.15	Cartoons (sequence)	Sequencing & Open-ended	5-point scale/3-point scale/not specified (sequencing); proportion of mental state terms in open-ended responses	Correctness; propensity
Reading the mind in the voice task	Golan et al., 2007	4	Population comparison (clinical); individual differences	Range: 19.3-35.6 Mean: 24.5	Audios	Forced-choice (4 options)	Binary scale	Correctness
Visual perspective taking task	Samson et al., 2010	4	Experimental condition comparison	Range: 21.7-40.9 Mean: 31.2	Pictorial probes	Forced-choice (binary)	Mean response time divided by proportion correct	Correctness
Comic strip task	Sarfati et al., 1997	3	Population comparison (clinical)	Range: 19.0-38.0 Mean: 27.4	Cartoons (sequence)	Forced-choice (3 options)	Binary scale	Correctness

Edinburgh Social Cognition Test (ESCoT)	Baksh et al., 2018	3	Individual differences	Range of means: 22.5-38.4 Mean of means: 32.8	Animations	Open-ended	4-point scale	Correctness
EmpaToM	Kanske et al., 2015	3	Neural underpinnings; individual differences	Range: 28.7-40.9 Mean: 36.8	Videos	Forced-choice (3 options)	Binary scale (score measure); n/a (RT measure)	Correctness
Moral judgment task	Young et al., 2007	3	Neural underpinnings	Range of means: 34.4-56.6 Mean of means: 41.7	Stories	Ratings	Rating differences between ToM and baseline conditions	Rating differences
Reading the mind in films task	Golan, Baron-Cohen, & Hill, et al., 2006	3	Population comparison (clinical); individual differences	Range: 35.6-38.4 Mean: 37.2	Videos	Forced-choice (4 options)	Binary scale	Correctness
Theory of mind stories task	Frith & Corcoran, 1996	3	Population comparison (clinical)	Range: 39-40.9 Mean: 39.6	Stories (with cartoons)	Open-ended	Binary scale	Correctness
Visual jokes test	Corcoran et al., 1997	3	Population comparison (clinical)	Range of means: 20.3-37.8 Mean of means: 27.0	Cartoons (single)	Open-ended	4-point scale/Binary scale	Correctness

Adult Theory of Mind test (A-ToM)	Brewer et al., 2017	2	Population comparison (clinical)	Range: 22.4-26.1 Mean: 24.3	Videos	Forced-choice (binary) & Open-ended	3-point scale/Binary scale;not applicable for RT	Correctness; RT
Attribution of intention task	Brunet, Sarfati, Hardy-Baylé & Decety, 2000	2	Neural underpinnings	Range: 30.9-47.7 Mean: 39.3	Cartoons (sequence)	Forced-choice (3 options)	Binary scale	Correctness
Cambridge mindreading face battery	Golan, Baron-Cohen & Hill, 2006	2	Population comparison (clinical)	Range: 22.2-22.5 Mean: 22.3	Videos	Forced-choice (4 options)	Binary scale	Correctness
Combined stories task	Achim et al., 2012	2	Population comparison (clinical)	Range: 24.2-25.2 Mean: 24.7	Stories	Open-ended	Binary scale/3-point scale	Correctness
False belief task (1st-order)	Wimmer & Perner, 1983	2	*Developmental differences (designed for children)	Range: 20.4-40.2 Mean: 30.3	Animations/ Cartoons (sequence)/Stories	Forced-choice (3 options)/Open-ended	Binary scale	Correctness
Mind Reading in Films task	Tahazadeh et al., 2020	2	Population comparison (clinical); individual differences	Range: 21.6-23.6 Mean: 22.6	Videos	Forced-choice (4 options)	Binary scale	Correctness
Modified Picture Stories-Theory of Mind Questionnaire (MPS-TOMQ)	Calso et al., 2019	2	Population comparison (age); individual differences	Range: 25.4-25.6 Mean: 25.5	Cartoons (sequence)	Sequencing & Open-ended	7-point scale (sequencing); n/a (sequencing time); not specified (TOMQ)	Correctness; n/a; not specified

Second-order false-belief task	Pickup & Frith, 2001	2	Population comparison (clinical)	Range: 32.7-33.5 Mean: 33.1	Playmobil figures/ Stories	Open-ended	3-point scale/4-point scale	Correctness
Situational test of emotion understanding	MacCann & Roberts, 2008	2	Individual differences	Range: 20.3-20.4 Mean: 20.4	Sentences	Forced-choice (5 options)	5-point scale	Not specified
Spontaneous ToM Protocol (STOMP)	Rice & Redcay, 2015	2	Neural underpinnings; individual differences	Mean: 20.3	Videos	Open-ended	Proportion of internal state statements	Propensity
Story comprehension test	Channon & Crawford, 2000	2	Lesion study	Range: 19.4-20.2 Mean: 19.8	Stories	Open-ended	3-point scale/binary scale*	Correctness(*); propensity*
Unexpected outcomes test	Dyck et al., 2001	2	*Developmental differences; individual differences (designed for children)	Range: 19.5-36.6 Mean: 28.1	Stories	Open-ended	3-point scale	Correctness
Virtual assessment of mentalising ability (VAMA)	Canty et al., 2017	2	Individual differences	Range: 25.9-45.6 Mean: 35.8	Interactive game	Forced-choice (4 options)	3-point scale/Binary scale	Correctness
Arena of Emotions Tasks	Rosenblau et al., 2015	1	Population comparison (clinical); individual differences	Mean: 32.4	Videos	Forced-choice (4 options)	Binary scale	Correctness

Attitudinal subset (APT) of the Aprosodia Battery	Orbelo et al., 2005	1	Population comparison (age)	Mean: 34.8	Audios	Forced-choice (binary)	Binary scale	Correctness
Belief-desires task	Apperly et al., 2011	1	Population comparison (age); experimental condition comparison	Mean: 20.3	Sentences	Forced-choice (binary)	n/a (RT measure)	n/a
Cartoon Reading the mind in the eyes task	Atherton, G. & Cross, L., 2021	1	Individual differences	Mean: 21.9	Cartoons (single)	Forced-choice (4 options)	Binary scale	Correctness
Cartoon stories ToM paradigm	Kosmidis, 2011	1	Population comparison (clinical); individual differences	Mean: 37.4	Cartoons (sequence)	Forced-choice (binary)	Binary scale	Correctness
Computerised false-belief task	Wang et al., 2021	1	Experimental condition comparison; individual differences	Mean: 19.5	Cartoons (sequence)	Forced-choice (binary)	n/a (RT measure)	n/a
Conflicting beliefs and emotions task	Shaw et al., 2004	1	* Lesion study (age not mentioned)	Mean: 30.6	Stories	Open-ended	Binary scale	Correctness
Conversations and Insinuations task	Ouellet et al., 2010	1	Population comparison (clinical)	Mean: 23.1	Videos	Forced-choice (4 options)	Binary scale	Correctness

Dewey Social Stories Test	Dewey, 1991	1	Population comparison (clinical)	Mean: 34.8	Stories	Forced-choice (4 options)	4-point scale	Deviation from most common response
Emotion Attribution task	Blair & Cipolotti, 2000	1	Lesion study	Mean: 40.2	Stories	Open-ended	Binary scale	Correctness
Faces test (Adoplhs et al.)	Adoplhs et al., 2002	1	*Lesion study (age not mentioned)	Mean: 36.6	Photos (face)	Forced-choice (binary)	Binary scale	Correctness
Faces test (Baron-Cohen et al.)	Baron-Cohen et al., 1997	1	Population comparison (clinical)	Mean: 20.7	Photos (face)	Forced-choice (binary)	Binary scale	Correctness
Irony perception task	Langdon et al., 2002	1	Population comparison (clinical)	Mean: 20.0	Stories	Forced-choice (binary)	Binary scale	Correctness
Joke-appreciation task	Happé et al., 1999	1	*Population comparison (clinical) (designed for the elderly)	Mean: 32.0	Cartoons (single)	Open-ended	4-point scale	Correctness
Judgement of preference	Girardi, MacPherson, & Abraham, 2011	1	Population comparison (clinical); experimental condition comparison	Mean: 38.4	Illustrated items	Forced-choice (4 options)	Binary scale	Correctness

Multifaceted Empathy Test	Dziobek et al., 2007	1	Population comparison (clinical)	Mean not reported	Photos (real person in context)	Forced-choice (4 options)	Binary scale	Correctness
Nonverbal cartoon task	Gallagher et al., 2000	1	Neural underpinnings	Mean: 42.0	Cartoons (single)	Open-ended	Binary scale	Correctness
Novel wisdom/ToM task	Rakoczy, H. et al., 2018	1	Population comparison (age); individual differences	Mean: 24.3	Stories	Open-ended	3-point scale	Correctness
Perspective Taking Task	Gallant, C., & Good, D., 2020	1	Population comparison (group split by other variables); individual differences	Mean: 19.8	Stories	Ratings	Average ratings for correct responses	Ratings
Pragmatic language comprehension task	Koster-Hale, Dodell-Feder, Saze, unpublished	1	n/a	Mean: 20.3	Sentences	Forced-choice (binary)	Binary scale	Not specified
Rutherford stories task	Rutherford, 2004	1	Experimental condition comparison	Mean: 24.7	Stories	Forced-choice (binary)	Binary scale	Correctness

Sandbox task	Sommerville et al., 2010	1	Population comparison (age); experimental condition comparison; individual differences	Mean: 37.7	Stories	Pointing to a location within a continuous space	Distance away from first location to second location	Distance away
Self-referential mentalizing interview	Ballespi, S. et al., 2019	1	Individual differences	Mean: 21.1	Interview questions	Ratings	n/a	n/a
Social Attribution Task-Multiple Choice	Klin, 2000	1	Population comparison (clinical)	Mean: 32.0	Animations	Forced-choice (4 options)/*Open-ended (original article)	Binary scale/*7-point scale/*Proportion of using mental state terms	Correctness/propensity
Social Cognition Screen Questionnaire (ToM subscale)	Roberts et al., 2011	1	* Individual differences (designed for clinical patients)	Mean: 37.8	Stories	Forced-choice (binary)	Binary scale	Correctness
Social stories questionnaire	Lawson, Baron-Cohen & Wheetwright, 2004	1	Population comparison (clinical)	Mean: 20.1	Stories	Forced-choice (binary)	Binary scale	Correctness
Story-Based Empathy Task	Dodich, A. et al., 2015	1	Norm setting	Mean: 49.6	Cartoons (sequence)	Forced-choice (3 options)	Binary scale (accuracy); 5-point scale (equivalent score)	Correctness; deviance from median

Strange stories film task	Murray et al., 2017	1	Population comparison (clinical); individual differences	Mean: 32.5	Videos	Open-ended	3-point scale	Correctness
Strange stories task + ToM Stories task	*Licata, M. et al., 2016 (the study that used this combined measure)	1	* n/a (refer to the two separate measures)	Mean: 38.0	Stories	Open-ended	4-point scale *(0/0.5/1/2)	Correctness
The cartoon vignette	Sebastian et al., 2012	1	Neural underpinnings	Mean: 21.3	Cartoons (sequence)	Forced-choice (binary)	Binary scale	Correctness
The situational test of emotion management	MacCann & Roberts, 2008	1	Individual differences	Mean: 20.4	Hypothetical scenarios	Forced-choice (4 options)/*Ratings (original article)	Binary scale/Weighted score (forced-choice); *distance from expert ratings (ratings)	Correctness(*); distance from expert rating*
Theory of Mind Assessment Scale (Th.o.m.a.s.)	Bosco et al., 2009	1	Population comparison (clinical)	Mean: 40.7	Interview questions	Open-ended	5-point scale	Coherence, clearness and abundance of contextualised examples
Theory of mind in dialogue	Dwyer et al., 2020	1	Population comparison (clinical)	Mean: 40.9	Interview questions	Open-ended	Number of references to own and others' beliefs	Propensity

ToM stories task	German & Hehman, 2006	1	Population comparison (age); individual differences	Mean: 38.8	Stories	Forced-choice (binary)	Binary scale	Correctness
ToM task (false belief + faux pas)	Henry et al., 2011	1	Population comparison (clinical)	Mean: 43.7	Stories	Open-ended	Binary (9 max) (FB1 total score); 3-point scale (FB2); 3-point scale (faux pas)	Correctness
ToM videos task (belief reasoning task)	Apperly et al., 2004	1	Lesion study	Mean: 38.8	Videos	Forced-choice (binary)	Binary scale	Correctness
ToM videos test	Sullivan & Ruffman, 2004	1	Population comparison (age); individual differences	Mean: 36.1	Videos	Forced-choice (binary)	Binary scale	Correctness
ToM-HCAT	Aykan, S. & Nalcaci, E., 2018	1	Individual differences	Mean: 21.3	Cartoons (single)	Forced-choice (4 options)	Binary scale	Correctness
Verbal stories ToM paradigm	Kosmidis, 2011	1	Population comparison (clinical); individual differences	Mean: 37.4	Stories	Open-ended	3-point scale (hinting task stories); not specified (FB1, FB2, 1st order deception, 2nd order deception)	Correctness

RT refers to response time.

Table 2.2. Reliability evidence of the top 8 measures in alphabetical order (number of studies providing positive/mixed/negative evidence)

Measure name	**Number of studies**	**Internal consistency**	**Test-retest reliability**	**Factor structure**	**Interrater reliability**
Animations Task*	10	1/0/0			4/0/0
FPRT*	28	6/0/0		5/1/0	5/0/0
Hinting Task*	25	0/1/5	0/2/0	1/0/0	2/0/0
Imposing memory Test	11	1/1/0			
MASC	11	2/1/0		1/0/0	
RMET	149	22/16/7	3/1/0	2/3/0	1/1/0 #
Strange Stories Task*	33	3/3/0		1/0/0	16/0/0
ToM Picture Stories Task	12	1/0/1			

* Tested in open-ended format in at least one study.

# Not tested in open-ended format but had interrater reliability reported (thus not included in the main analysis or Figure 2).

Table 2.3. Validity evidence of the top 8 measures in alphabetical order (number of studies providing positive/mixed/negative evidence).

Measure name	Number of studies	Known-group validity	Criterion-related validity	(Broad) Convergent validity	Discriminant validity
Animations Task	10	4/0/1		0/0/2	3/1/1
FPRT	28	4/0/0	1/0/1	5/1/2 (2)	2/0/0
Hinting task	25	8/0/1	0/0/1	6/1/2	3/0/0
Imposing memory test	11		2/0/0	3/2/1	2/0/0
MASC	11	2/0/0		2/0/0	1/0/0
RMET	149	14/0/1	1/0/1	27/7/11 (1)	0/0/1
Strange Stories Task	33	10/0/1	0/0/2	4/0/1(1)	7/0/1
ToM Picture Stories Task	12	3/0/1		2/1/1	

The number of studies that report a relevant significance test without specifying the effect size is marked in parentheses, if applicable.

Table 2.4. POMP score of the 75 identified measures (in alphabetical order). Measures showing evidence of ceiling effects are highlighted in red. The "top eight" most frequently-used measures discussed in most detail in the text are shaded in gray.

Measure name	Stimulus type	Number of studies	Mean POMP score for neurotypical adults	POMP score range
Adult Theory of Mind test (A-ToM)	Videos	2	87.25%	n/a
Animations task	Animations	10	Appropriateness: 64.85% (7 studies) Feelings: 51.76% (2 studies) Intentionality: 66.2% (1 study)	Appropriateness: 41.13%-75.75% Feelings: 49.13%-54.38% Intentionality: n/a
Arena of Emotions Tasks	Videos	1	Indirect: 68% Direct: 67%	n/a
Attitudinal subset (APT) of the Aprosodia Battery	Audios	1	Not reported	n/a
Attribution of intention task	Cartoons (sequence)	2	84.43% (1 study)	n/a
Belief-desires task	Sentences	1	n/a	n/a
Cambridge mindreading face battery	Videos	2	75.59%	72.00%-79.18%
Cartoon Reading the mind in the eyes task	Cartoons (single)	1	67.00%	n/a
Cartoon stories ToM paradigm	Cartoons (sequence)	1	82.41%	n/a

Combined stories task	Stories	2	1st order: 93.33% (1 study) 2nd order: 83.85% (1 study)	n/a
Comic strip task	Cartoons (sequence)	3	88.80% (2 studies)	82.96%-94.64%
Computerised false-belief task	Cartoons (sequence)	1	n/a	n/a
Conflicting beliefs and emotions task	Stories	1	1st order belief: 98.00% 2nd order belief: 96.50% 1st order emotion: 89.25% 2nd order emotion: 92.50%	n/a
Conversations and Insinuations task	Videos	1	73.80%	n/a
Dewey Social Stories Test	Stories	1	92.42%	n/a
Director task	Interactive game	4	Ambiguous experimental trials: 96.80% (2 studies) Relational experimental trials: 58.00% (1 study)	Ambiguous trials: 95.00%-98.60%
Edinburgh Social Cognition Test (ESCoT)	Animations	3	Cognitive ToM: 74.18% (2 studies) Affective ToM: 88.18% (2 studies)	Cognitive ToM: 73.00%-75.37% Affective ToM: 86.93%-89.43%
Emotion Attribution task	Stories	1	90.43%	n/a
EmpaToM	Videos	3	80.48% (2 studies)	71.61%-89.35% (2 studies)
Faces test (Adoplhs et al.)	Photos (face)	1	Not reported	n/a

Faces test (Baron-Cohen et al.)	Photos (face)	1	Not reported	n/a
False belief task (1st-order + 2nd-order)	Cartoons (sequence)/Stories	8	1st + 2nd order: 91.12% (3 studies) 1st order: 90.77% (3 studies) 2nd order: 73.46% (3 studies)	1st + 2nd order: 84.89%-94.99% (3 studies) 1st order: 86.30%-95.00% (3 studies) 2nd order: 65.00%-89.57% (3 studies)
False belief task (1st-order)	Animations/Cartoons (sequence)/Stories	2	87.97%	75.93%-100%
Faux pas recognition test	Stories	28	85.90% (20 studies)	69.90%-96.00% (20 studies)
Hinting task	Stories/Videos	25	81.31% (21 studies)	62.19%-93.05% (21 studies)
Imposing memory test	Stories/Videos	11	82.44% (5 studies)	74.40%-84.13% (5 studies)
Irony perception task	Stories	1	Hit: 78.00% False alarm: 20.00% Sensitivity: 87.00%	n/a
Joke-appreciation task	Cartoons (single)	1	55.33%	n/a
Judgement of preference	Illustrated items	1	Not reported	n/a
MASC	Videos	11	Total correct: 73.57% (8 studies) Cognitive: 77.77% (2 studies) Affective: 76.45% (2 studies)	Total correct: 59.09%-78.42% (8 studies) Cognitive: 76.65%-78.89% (2 studies) Affective: 75.56%-77.33% (2 studies)
Mind Reading in Films task	Videos	2	64.89%	59.96%-69.81%

Modified Picture Stories-Theory of Mind Questionnaire (MPS-TOMQ)	Cartoons (sequence)	2	MPS: 85.81% (1 study) TOMQ: 55.82%	MPS: n/a TOMQ: 44.64%-67.00%
Moral judgment task	Stories	3	n/a	n/a
Multifaceted Empathy Test	Photos (real person in context)	1	Not reported	n/a
Nonverbal cartoon task	Cartoons (single)	1	97.27%	n/a
Novel wisdom/ToM task	Stories	1	90.90%	n/a
Perspective Taking Task	Stories	1	n/a	n/a
Picture sequencing task	Cartoons (sequence)	4	86.39%	82.33%-92.00% (3 studies)
Pragmatic language comprehension task	Sentences	1	Pragmatic inference accuracy: 81.90%	n/a
Reading the mind in films task	Videos	3	64.09% (1 study)	n/a
Reading the Mind in the Eyes Test	Photos (eyes)	149	Total: 72.00% (125 studies) Positive: 70.73% (7 studies) Neutral: 69.89% (7 studies) Negative: 71.36% (7 studies)	Total: 57.84%-86.12% (125 studies) Positive: 64.92%-82.00% (7 studies) Neutral: 62.50%-75.00% (7 studies) Negative: 60.00%-85.72% (7 studies)
Reading the mind in the voice task	Audios	4	71.00% (3 studies)	64.00%-78.00% (3 studies)

Rutherford stories task	Stories	1	Unweighted score: 90.00%	n/a
Sandbox task	Stories	1	n/a	n/a
Second-order false-belief task	Playmobil figures/Stories	2	57.75% (1 study)	n/a
Self-referential mentalizing interview	Interview questions	1	n/a	n/a
Short Story Task	Stories	5	Mental state reasoning: 50.17% (3 studies) Total: 63.71% (2 studies) Spontaneous mental state reasoning: 19.00% (1 study)	Mental state reasoning: 38.69%-58.06% (3 studies) Total: 59.22%-68.19% (2 studies)
Situational test of emotion understanding	Sentences	2	Not available	n/a
Social Attribution Task-Multiple Choice	Animations	1	80.95%	n/a
Social Cognition Screen Questionnaire (ToM subscale)	Stories	1	84.30%	n/a
Social stories questionnaire	Stories	1	Subtle utterances: 29.10% Blatant utterances: 57.50% Non-existence utterances: 92.15%	n/a
Spontaneous ToM Protocol (STOMP)	Videos	2	30.11%	29.11%-39.10%
Story comprehension test	Stories	2	65.50%	65.00%-66.00%

Story-Based Empathy Task	Cartoons (sequence)	1	Total: 87.39% Intention attribution: 89.33% Emotion attribution: 87.00%	n/a
Strange stories film task	Videos	1	Intention: 80.21% Mental state talk: 49.38% Interaction: 72.71%	n/a
Strange Stories Task	Stories	33	87.37% (25 studies)	55.00%-99.50% (25 studies)
Strange stories task + ToM Stories task	Stories	1	63.85%	n/a
TASIT	Videos	8	Part 2: 88.68% (4 studies) Part 3: 84.87% (7 studies)	Part 2: 84.42%-91.80% (4 studies) Part 3: 83.20%-86.70% (7 studies)
The cartoon vignette	Cartoons (sequence)	1	Affective ToM: 86.50% Cognitive ToM: 91.94%	n/a
The situational test of emotion management	Hypothetical scenarios	1	Not reported	n/a
Theory of Mind Assessment Scale (Th.o.m.a.s.)	Interview questions	1	First-person ToM: 95.50% Third-person allocentric ToM: 92.50% Third-person egocentric: 92.75% Second-order ToM: 91.50%	n/a
Theory of mind in dialogue	Interview questions	1	n/a	n/a
Theory of mind stories task	Stories (with cartoons)	3	Total: 90.15% (1 study)	n/a

ToM Picture Stories task	Cartoons (sequence)	12	Total: 91.63% (6 studies) Sequencing: 86.94% (5 studies) Questionnaire: 92.45% (5 studies)	Total: 89.42%-94.34% (6 studies) Sequencing: 70.00%-94.44% (5 studies) Questionnaire: 81.86%-95.83% (5 studies)
ToM stories task	Stories	1	75.29%	n/a
ToM task (false belief + faux pas)	Stories	1	n/a	n/a
ToM videos task (belief reasoning task)	Videos	1	87.39%	n/a
ToM videos test	Videos	1	88.08%	n/a
ToM-HCAT	Cartoons (single)	1	70.72%	n/a
Unexpected outcomes test	Stories	2	60.75% (1 study)	n/a
Verbal stories ToM paradigm	Stories	1	Hinting: 92.17% 1st order false belief: 97.50% 2nd order false belief: 80.00% 1st order deception: 96.00% 2nd order deception: 90.00%	n/a
Virtual assessment of mentalising ability (VAMA)	Interactive game	2	Cognitive: 66.68% (frequency); 72.65% (cumulative; 1 study) Affective: 61.50% (frequency); 69.93% (cumulative; 1 study) Total: 62.50% (frequency; 1 study)	Cognitive: 64.35%-69.00% (frequency) Affective: 60.65%-62.35% (frequency)

Visual jokes test	Cartoons (single)	3	58.00%	55.00%-66.25%
Visual perspective taking task	Pictorial probes	4	n/a	n/a
Yoni task	Illustrated items	6	Total: 92.86% (1 study) Affective: 89.62% (3 studies) Cognitive: 87.33% (3 studies)	Affective: 84.35%-92.55% Cognitive: 83.10%-90.44%

Table 2.5. Interrelations among identified ToM measures (in alphabetical order). Tasks that did not show any correlation with other measures with an effect size larger than the .19 threshold are highlighted in red.

Task name	Correlated task	Number of studies	Correlation index range	Number of studies reporting $r \geq .19$ (n/a)	Number of studies reporting significant correlation
Adult Theory of Mind test (A-ToM)	Animations task	1	.12-.17	0	0
	Strange Stories Task	1	.50	1	1
Animations task	Adult Theory of Mind test (A-ToM)	1	.12-.17	0	0
Arena of Emotions Tasks	RMET	1	.303-.417	1	1
Belief-desires task	Imposing memory test	1	.048	0	0
	Pragmatic language comprehension task	1	.056	0	0
	RMET	1	.115	0	0
	Spontaneous ToM Protocol (STOMP)	1	-.023	0	0
Cartoon stories ToM paradigm	Verbal stories ToM paradigm	1	.008-.529	1	1
Combined stories task	Comic strip task	1	.08	0	0

Comic strip task	Combined stories task	1	.08	0	0
Director task	Visual perspective taking task	1	-.18	0	1
Dewey Social Stories Test	Faux pas recognition test	1	-.276	1	1
	RMET	1	-.143	0	0
Edinburgh Social Cognition Test (ESCoT)	Judgement of preference	1	not reported	0 (1)	0
	Reading the mind in films task	1	.36-.42	1	1
	RMET	2	.25-.48	2	2
	Visual perspective taking task	1	-.07 - -.34	1	1
Emotion Attribution task	RMET	1	.43	1	1
	Strange Stories Task	1	.69	1	1

	ToM Picture Stories task	1	.46	1	1
EmpaToM	Visual perspective taking task	1	.17	0	1
Faces test (Baron-Cohen et al.)	RMET	1	.29	1	1
	Reading the mind in the voice task	1	.22	1	1
False belief task (1st-order + 2nd-order)	RMET	1	.12	0	0
False belief task (1st-order)	RMET	1	.12	0	0
Faux pas recognition test	Dewey Social Stories Test	1	-.276	1	1
	RMET	5	.13-.407	4	4
	Strange Stories Task	2	.11; not reported	0 (1)	0
	ToM Picture Stories task	1	.18	0	1
	Virtual assessment of mentalising ability (VAMA)	1	.04-.45	1	1

Hinting task	Imposing memory test	2	.21	2	2
	RMET	3	.097-.28	2	2
	Second-order false-belief task	1	.201-.276	1	1
	Situational test of emotion understanding	2	.30-.33	2	2
	Social Attribution Task-Multiple Choice	1	.117	1	0
	TASIT	1	.25	1	1
	The situational test of emotion management	1	.22	1	1
	ToM Picture Stories task	1	.146	0	0
	Virtual assessment of mentalising ability (VAMA)	1	.05-.36	1	1
	Visual jokes test	1	Kendall’s tau=.05 (transformed r=0.078 (Gilpin, 1993))	1	0

Imposing memory test	Belief-desires task	1	.048	1	0
	Hinting task	2	.21	2	2
	Pragmatic language comprehension task	1	-.051	1	1
	RMET	6	-.069-.42	4	4
	Situational test of emotion understanding	2	.44-.48	2	2
	Spontaneous ToM Protocol (STOMP)	2	.125-.28	1	1
	The situational test of emotion management	1	.39	1	1
Judgement of preference	Edinburgh Social Cognition Test (ESCoT)	1	not reported	0 (1)	0
	Reading the mind in films task	1	not reported	0 (1)	0
	RMET	1	not reported	0 (1)	0
MASC	RMET	1	.30	1	1
	Self-referential mentalizing interview	1	not reported; .25	1	1

Mind Reading in Films task	RMET	1	.56	1	1
Perspective Taking Task	RMET	1	-.007-.256	1	1
Picture sequencing task	Theory of mind stories task	1	.55-.63	1	1
Pragmatic language comprehension task	Belief-desires task	1	.056	0	0
	Imposing memory test	1	-.051	0	0
	RMET	1	.068	0	0
	Spontaneous ToM Protocol (STOMP)	1	.015	0	0
Reading the mind in films task	Edinburgh Social Cognition Test (ESCoT)	1	.36-.42	1	1

	Judgement of preference	1	not reported	0 (1)	0
	RMET	2	.38-.62	2	2
RMET	Arena of Emotions Tasks	1	.303-.417	1	1
	Belief-desires task	1	.115	0	0
	Dewey Social Stories Test	1	-.143	0	0
	Emotion Attribution task	1	.43	1	1
	Edinburgh Social Cognition Test (ESCoT)	2	.25-.48	2	2
	Faces test (Baron-Cohen et al.)	1	.29	1	1
	False belief task (1st-order + 2nd-order)	1	.12	0	0
	False belief task (1st-order)	1	.12	0	0
	Faux pas recognition test	5	.13-.407	4	4

	Hinting task	3	.097-.28	2	2
	Imposing memory test	6	-.069-.42	4	4
	Judgement of preference	1	not reported	0 (1)	0
	MASC	1	.30	1	1
	Mind Reading in Films task (Tahazadeh et al.)	1	.56	1	1
	Perspective Taking Task (scenarios from Hynes et al.)	1	-.007-.256	1	1
	Pragmatic language comprehension task	1	.068	0	0
	Reading the mind in films task	2	.38-.62	2	2
	Reading the mind in the voice task	1	.35	1	1
	Short Story Task (Dodell-Feder et al.)	4	.18-.42	3	4
	Situational test of emotion understanding	2	.53-.54	2	2

	Social Attribution Task-Multiple Choice	1	.331	1	1
	Spontaneous ToM Protocol (STOMP)	2	-.16 - -.115	0	0
	Strange Stories Task	4	.14-.42; not reported	2 (1)	1
	TASIT	1	.371	1	1
	The situational test of emotion management	1	.42	1	1
	ToM Picture Stories task	2	.43-.535	2	2
	Unexpected outcomes test	1	.26	1	1
	Yoni task	1	.26	1	1
Reading the mind in the voice task	Faces test (Baron-Cohen et al.)	1	.22	1	1
	RMET	1	.35	1	1
Second-order false-belief task	Hinting task	1	.201-.276	1	1
Self-referential mentalizing interview	MASC	1	not reported; .25	1	1
Short Story Task	RMET	4	.18-.42	3	4

Situational test of emotion understanding	Hinting task	2	.30-.33	2	2
	Imposing memory test	2	.44-.48	2	2
	RMET	2	.53-.54	2	2
	The situational test of emotion management	1	.62	1	1
Social Attribution Task-Multiple Choice	Hinting task	1	.117	0	0
	RMET	1	.331	1	1
Spontaneous ToM Protocol (STOMP)	Belief-desires task	1	-.023	0	0
	Imposing memory test	2	.125 - .28	1	1
	Pragmatic language comprehension task	1	.015	0	0
	RMET	2	-.16 - -.115	0	0
Strange Stories Task	Adult Theory of Mind test (A-ToM)	1	.50	1	1

	Emotion Attribution task	1	.69	1	1
	ToM Picture Stories task	1	.42	1	1
	Faux pas recognition test	2	.11; not reported	0 (1)	0
	RMET	4	.14-.42; not reported	2 (1)	1
TASIT	Hinting task	1	.25	1	1
	RMET	1	.371	1	1
	ToM Picture Stories task	1	.525	1	1
The situational test of emotion management	Hinting task	1	.22	1	1

	Imposing memory test	1	.39	1	1
	RMET	1	.42	1	1
	Situational test of emotion understanding	1	.62	1	1
Theory of mind stories task	Picture sequencing task	1	.55-.63	1	1
ToM Picture Stories task	Emotion Attribution task	1	.46	1	1
	Faux pas recognition test	1	.18	0	1
	Hinting task	1	.146	0	0
	RMET	2	.43-.535	2	2
	Strange Stories Task	1	.42	1	1
	TASIT	1	.525	1	1
Unexpected outcomes test	RMET	1	.26	1	1
Verbal stories ToM paradigm	Cartoon stories ToM paradigm	1	.008-.529	1	1
Virtual assessment of mentalising ability (VAMA)	Faux pas recognition test	1	.04-.45	1	1

	Hinting task	1	.05-.36	1	1
	Yoni task	1	.01-.21	1	0
Visual jokes test	Hinting task	1	Kendall’s tau=.05 (transformed r=.078 (Gilpin, 1993))	0	0
Visual perspective taking task	Director task	1	-.18	0	1
	Edinburgh Social Cognition Test (ESCoT)	1	- .34 - -.07	1	1
	EmpaToM	1	.17	0	1
Yoni task	RMET	1	.26	1	1
	Virtual assessment of mentalising ability (VAMA)	1	.01-.21	1	0

Table 2.6. Average Cronbach's alpha of the top 8 measures (in alphabetical order).

Task name	Average Cronbach's alpha	Number of reports
Animations task	0.8	1
Faux pas recognition test	0.87	7
Hinting task	0.55	6
Imposing memory test	0.86	1
MASC	0.76	3
Reading the Mind in the Eyes Test	0.68	37
Strange Stories Task	0.68	5
ToM Picture Stories task	0.65	2

Table 2.7. Interrelations among top 8 measures before and after correction for attenuation (in alphabetical order).

Task	Correlated task	Range of *r*	Range of corrected *r*	Number of studies with uncorrected *r*≥.19 (n/a)	Number of studies with corrected *r*≥.19 (n/a)	Number of reports
Faux pas recognition test	Reading the Mind in the Eyes Test	.13-.41	.17 - .53	4	4	5
	Strange Stories Task	.11; not reported	.14; not reported	0 (1)	0 (1)	2
	ToM Picture Stories task	.18	.24	0	1	1
Hinting task	Imposing memory test	.21	.31	2	2	2
	Reading the Mind in the Eyes Test	.10 - .28	.16 - .46	2	2	3
	ToM Picture Stories task	.15	.25	0	1	1
Imposing memory test	Hinting task	.21	.31	2	2	2
	Reading the Mind in the Eyes Test	-.07 - .42	-.09 - .55	4	4	7

MASC	Reading the Mind in the Eyes Test	.30	.42	1	1	1
Reading the Mind in the Eyes Test	Faux pas recognition test	.13 - .407	.17 - .53	4	4	5
	Hinting task	.10 - .28	.16 - .46	2	2	3
	Imposing memory test	-.07 - .42	-.09 - .55	4	4	7
	MASC	.30	.42	1	1	1
	Strange Stories Task	.14 - .42; not reported	.21 - .62; not reported	2 (1)	3 (1)	4
	ToM Picture Stories task	.43 - .54	.65 - .81	2	2	2
Strange Stories Task	Faux pas recognition test	.11; not reported	.14; not reported	0 (1)	0 (1)	2
	Reading the Mind in the Eyes Test	.14 - .42; not reported	.21 - .62; not reported	2 (1)	3 (1)	4

	ToM Picture Stories task	.42	.63	1	1	1
ToM Picture Stories task	Faux pas recognition test	.18	.24	0	1	1
	Hinting task	.15	.25	0	1	1
	Reading the Mind in the Eyes Test	.43 - .54	.65 - .81	2	2	2
	Strange Stories Task	.42	.63	1	1	1

Chapter 3

Do mindreading interpretations differ between age groups?

3.1 Introduction

Mindreading involves the attribution of mental states to others. As identified in Chapter 2, individual differences in mindreading are most frequently examined in terms of correctness or accuracy. The term "accuracy" presupposes a "correct" answer, but existing tasks have been criticised as lacking an objective and observable ground truth as the basis for considering the "accurate" answer (Long et al., 2022). This issue is especially crucial to tasks that involve the interpretation of social stimuli and do not provide a non-arbitrary criterion for correctness, for example, the Faces test (Adoplhs et al., 2002; Baron-Cohen et al., 1997) or the Reading the Mind in the Eyes Test (RMET; Baron-Cohen et al., 2001). In these tasks, even the creators have no access to what the people featured in the stimuli were thinking or feeling. The correct answers are usually based on the expert opinion of researchers or results from pilot studies with a small number of participants. It is therefore unclear whether the "correct" answers represent the most appropriate interpretation. One way to address this problem is to develop tasks that have a pre-established ground truth, like the Empathic Accuracy task (Ickes et al., 1986; Ickes, 1993) or the Interviews Task (Long et al., 2022). However, given research showing that performance in the empathic accuracy task is more related to the expressivity of the target (Zaki et al., 2008) and one's familiarity with the target (Zaki et al., 2009) than trait empathy, evaluating participants' mindreading ability by comparing their responses to the actual thoughts or feelings of the targets may not always be the most promising approach.

An alternative to the ground truth approach is to use social agreement as a criterion for characterising mindreading success. While this idea is not new, since some tasks already use expert or pilot participant consensus for "correct" answers as mentioned above, the proposed approach does not make claims about actual "accuracy" or assume that social agreement leads to only one successful answer. In other words, an overarching aim of the current chapter is to provide an alternative perspective to the notion of “accuracy” in such mindreading tasks, by showing the presence of multiple legitimate interpretations and variation in endorsement of interpretations by different groups of neurotypical adults, with a series of three empirical studies.

In tasks adopting stimuli without a known ground truth, it is questionable to regard test takers as worse mindreaders if they have not selected the “correct” answers, as it is possible that there are multiple interpretations that differ among individuals. In the clinical literature, the common mental state interpretation endorsed by the neurotypical adult population is generally regarded as the correct answer, setting the benchmark for comparing clinical groups’ responses. Deviations by clinical groups from the neurotypical norm are considered errors. For example, “over-mentalising” errors and “under-mentalising” errors that deviate from the assumed correct answer in the MASC are commonly found in people with schizophrenia (Peyroux et al., 2019; Sharp & Hernandez, 2021). In examining mindreading performance in clinical groups, there is a basis to argue for these alternative interpretations being errors because they correlate with positive and negative symptoms, respectively (e.g., Peyroux et al., 2019). Some theories further consider mindreading deficits as an explanation for differences in social and occupational functioning in neurodivergent populations, especially autistic people, represented by the mind-blindness theory (Baron-Cohen, 1995; Baron-Cohen et al., 1985). However, research on the double empathy problem between neurotypical and autistic individuals has suggested that not only do autistic individuals find interpreting the mental states of their neurotypical counterparts difficult, but neurotypical individuals also find difficulty interpreting the mental states of autistic individuals (Alkhaldi et al., 2019; Edey et al., 2016; Milton, 2012), which questions whether neurotypical individuals always have the authority to determine the correct interpretation when it comes to attribution of mental states to clinical populations.

Furthermore, even assuming consensus among neurotypical individuals as a proxy for ground truth, neurotypical individuals might vary in their attribution of mental states to others in the same scenario, leading to multiple legitimate interpretations. This can undermine the validity of assuming a single "correct" answer. Hence, the question of whether there are multiple common interpretations among neurotypical individuals warrants more empirical investigation.

A further problem lies in the conceptualisation of "accuracy": consensus among group members, or the most common interpretation, is not necessarily the most accurate answer. This is analogous to how the most popular answer in an intelligence test item might not be the correct answer, especially for difficult items designed so that only a small proportion of participants can obtain the correct answer. The notion of "accuracy" is therefore dubious concerning mindreading tasks that feature social stimuli without a known ground truth or authorial intention. With reference to using the common interpretations within neurotypical adults as the baseline for establishing "accuracy", the criterion is more focused on how well people align with other people in terms of their mental state interpretations of others. Subsequently, the term "alignment", adapted from literature on social coordination (Perez-Zapata & Apperly, 2022), is suggested as an alternative term for characterising success in such tasks.

Inspired by Perez-Zapata and Apperly (2022) who calculated alignment scores to measure similarity of a participant's responses with other participants in pure coordination games, in the current studies, alignment is assessed by directly comparing an individual's interpretation with that of others and calculating how many people agree with the interpretation. Variations in alignment could indicate individual differences in mindreading performance. Furthermore, we expect that people's alignment with others varies with the comparison baselines formed by people who possess different features. For example, culture has been shown to influence mindreading performance: people found it easier to reason about the mental states of those who they perceived to be from the same cultural group (Adams et al., 2010; Perez-Zapata et al., 2016). This effect also was observed in perceived familiarity with the agents (Zaki et al., 2009). These findings suggest that people might align with similar individuals more than dissimilar individuals. If an individual aligns well not only with similar peers but also with dissimilar counterparts, it may characterise superior flexibility in mindreading.

To summarise, it is problematic to assume that mindreading “accuracy” is captured by comparing participants’ answers to the predetermined correct answers, and the notion of “accuracy” itself is problematic in the case of interpreting social stimuli without a known ground truth. The studies in the current chapter address two questions. First, is there only one interpretation that is considered the most appropriate or plausible among neurotypical individuals, or are there multiple interpretations that are perceived to be highly probable? Second, do neurotypical individuals differ from each other in their perceived best interpretations? One further question is related to individual differences: do people show consistent individual differences in their alignment with others across scenarios?

If variation in interpretation is observed across any dimension, such as age, gender, or ethnicity, it provides evidence for individual differences in mental state interpretations. In the current studies the focus is on age because older adults may have accumulated more diverse social experiences over their lifetimes than younger adults leading to more social insight (Happé et al., 1998), which may drive differences in mental state interpretations between the two age groups. Studies 1 and 2 were pilot studies for the more extensive study 3. Study 1 explored whether younger adults and older adults shared common interpretations of what an agent was thinking or feeling in ambiguous social scenarios. Building on preliminary findings from study 1, study 2 compared participants’ alignment with members of their own age group and the other age group in mental state interpretations with a fine-grained coding scheme generated with a bottom-up approach. The same coding scheme was adopted in Study 3 that recruited a large sample, in which the number of popular interpretations for each item was compared to simulated baselines and inter-item correlations were investigated to answer the question on individual differences in mindreading. It was predicted that there would be multiple interpretations of the same ambiguous social stimulus and the endorsement of these interpretations would vary between the two age groups.

3.2 Study 1: Pilot study

Study 1 examined differences in mental state interpretations of pictorial stimuli featuring social scenarios without a known ground truth between younger adults and older adults. This exploratory pilot study aimed to (1) explore whether participants generated diverse interpretations of the same social stimulus and (2) whether these interpretations differed between the two age groups.

3.2.1 Method

Participants. Twelve younger (18-25 years; Mage = 21.75) and 12 older adults (53-60 years; Mage = 57.33) with balanced gender were recruited online via Prolific with the following screening criteria: UK residence, speaking English as their first language, and had not been diagnosed with ASD. All participants received £3.75 for completing the study. The study was approved by the Ethical Review Committee at the University of Birmingham. The study was not preregistered.

Materials. Ten coloured photos depicting two target adults engaging in social interactions were presented (refer to figure 3.1 for two examples). The target adults depicted across stimuli varied in sex, age, and ethnicity. These photos originated from pilot work conducted in the laboratory associated with the research team (Yeung, Apperly). This pilot work involved searching the Internet for Creative Commons-licensed images depicting individuals in social interactions, which could have included genuine interactions or acted scenarios. These stimuli had been adopted in previous studies on social coordination and mindreading (Perez-Zapata, 2023; Pomareda, 2023).

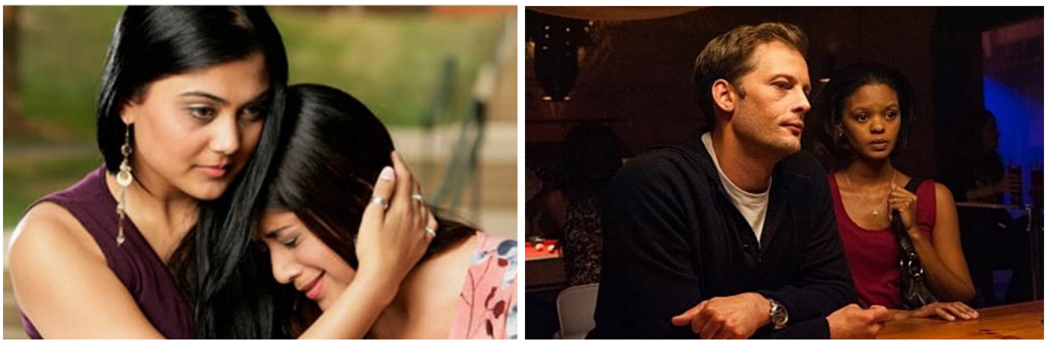


Figure 3.1. Two examples of pictorial stimuli presented.

Design and procedure. Within each age group, each participant's interpretations of each picture were compared with the most popular interpretation(s) with others from their age group, or from the other age group, based on their responses in a Qualtrics questionnaire. Each participant saw five of the 10 pictures in a randomised order. First, the participants were required to describe what they thought was happening in each picture in the first block of questions (the description block). Then the participants were presented the same picture again, but in this block (the interpretation block) one of the characters was circled in each picture (i.e., the mindreading target). With each picture presented, participants were asked the question "What do you think the circled person is thinking/feeling?" and they entered open-ended responses. Multiple interpretations were allowed, but participants were asked to enter only one interpretation a time. This was achieved by showing a question "Can you think of any other possibilities?" every time a participant submitted one interpretation. If the participant selected "yes", they were redirected to input their second response for the same picture. A maximum of 15 different interpretations was allowed for each picture. The total duration of the task was around 30 minutes.

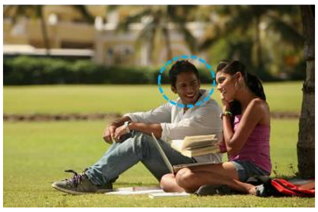
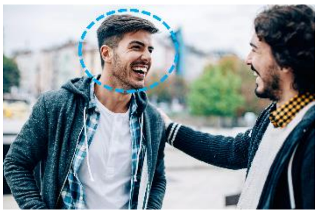
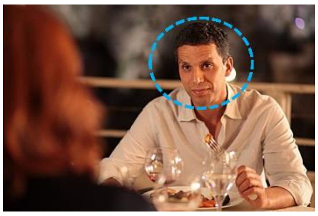
3.2.2 Results and discussion

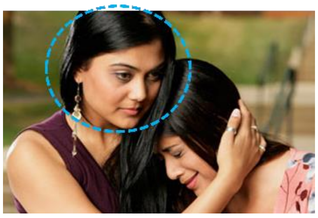
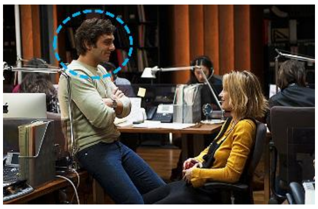
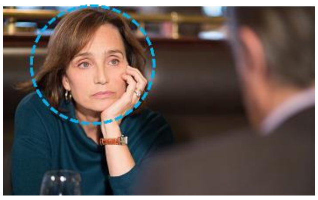
No formal statistical test was involved in Study 1 due to its small sample size and its nature as a preliminary pilot study. Responses in the description block were mostly descriptions of the environment, so only the responses in the interpretation block are reported below.

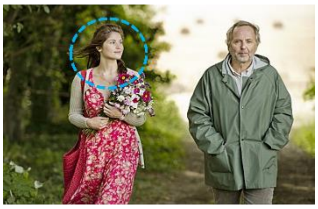
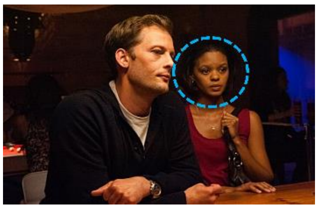
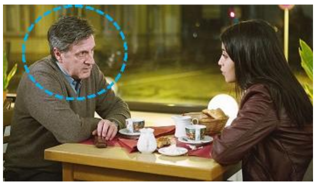
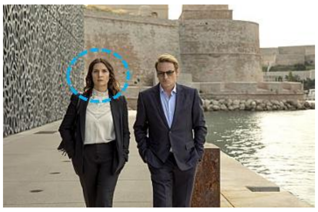
The text responses were qualitatively inspected and grouped into categories. Table 3.1 tabulates the most popular categories of interpretations to indicate any differences between the two age groups. The most popular category/ies for each item was determined by the number of participants who had generated responses that involved the category/ies for the picture (i.e., item) concerned. If a single category was involved in the responses by 50% or more participants, it was selected as the most popular category and reported below. If the proportion of endorsement of the most popular category was below 50%, the second most common category/ies were also included and reported below. The label of each item always starts with a “P”, which stands for “Picture”, followed by an index number from 1 to 10.

Except for P2 and P7, the two age groups did not show the same combination of the most frequent categories for any item. When the same category overlapped between groups, the proportion of participants endorsing it often differed. This difference in proportion of endorsement will be discussed in a more quantitative method in Studies 2 and 3.

Table 3.1. The most popular categories in the two age groups for each item in Study 1.

Picture (P stands for Picture)	Age group	
	Old	**Young**
P1 	• Romantic feelings (58.33%) (e.g. "He is wishing that he was her boyfriend")	• Romantic feelings (38.46%) (e.g. "love towards the girl"); • Happy (30.77%) (e.g. "The person is feeling happy with the choice to study with this girl")
P2 	• Happy (81.82%) (e.g. "He is thinking he is very pleased to see his friend.")	• Happy (88.89%) (e.g. "They are feeling happy about seeing their friend")
P3 	• Negative experience (31.25%) (e.g. "He is hurt"); • Surprised (31.25%) (e.g. "I think he's feeling surprised")	• Romantic feelings (35.71%) (e.g. "intent on being with this woman"); • Happy (14.29%) (e.g. "This person is feeling content."); • Intrigued (14.29%) (e.g. "interested into what is being said by the lady")

P4 	• Compassion (46.67%) (e.g. “She is feeling emotion and sadness for her friend/sister”); • Sinister intent (20%) (e.g. “Avarice”)	• Compassion (38.46%) (e.g. This woman is feeling sorrow for the person next to her who is crying.); • Embarrassed (15.38%) (e.g. “This person feels uncomfortable/awkward due to the current situation she is in.”)
P5 	• Romantic feelings (53.33%) (e.g. “I cannot keep my eyes off her.”)	• Romantic feelings (41.67%) (e.g. ” He is attracted to the woman sat down”); • Happy (41.67%) (e.g. “satisfaction that this is the situation he has ended up in”)
P6 	• Bored (78.57%) (e.g. “Bored”)	• Bored (38.46%) (e.g. “This person is bored of the conversation.”); • Attentive (23.08%) (e.g. “The person is listening intently.”; • Distressed (23.08%) (e.g. “distressed”)

P7 	• Happy (75%) (e.g. "content")	• Happy (75%) (e.g. "The person is feeling happy and having a good day.")
P8 	• Surprised/shocked (46.15%) (e.g. "Surprise"); • Anxious (23.08%) (e.g. "She feels nervous")	• Anxious (40%) (e.g. "This person is waiting for the bartender and looks anxious."); • Surprised/shocked (30%) (e.g. "startled")
P9 	• Concerned (33.33%) (e.g. "Concerned and worried"); • Angry/irritated (25%) (e.g. "He is angry/annoyed")	• Concerned (20%) (e.g. "concerned"); • Stressed (20%) (e.g. "He looks stressed out and serious.")
P10 	• Negative emotion (100%) (e.g. "Supressed anger or irritation")	• Determined (40%) (e.g. "I think she is feeling determined."); • Negative emotion (30%) (e.g. "Disgust")

In summary, this pilot study suggested that (1) individuals varied in how they interpreted the mental states of the targets and (2) there were signs of variation between the two age groups. The average number of unique interpretations (that were coded as different categories) generated by each participant across the five stimuli ranged from 1 to 2.4 (mean across participants = 1.57).

Six (P3, P4, P5, P8, P9 and P10) out of the 10 pictures were selected as stimuli for the later studies to shorten the task. The stimuli were chosen to diversify the potential interpretations generated across items. For example, “happy” and “romantic feelings” were the most popular categories in P1, P2 and P5, so P5 has been kept to contain an item that is likely to generate interpretations with a positive valence. The other items were selected based on the principle that the proportion of endorsement of the most popular category did not exceed 50% in either group, except for P10, as participants’ responses to P10 could be possibly coded into more fine-grained categories of negative emotions in the subsequent studies with a larger sample size.

3.3 Study 2: Construction of coding scheme

Results of study 1 suggested that older and younger adults could provide different mental state interpretations about the same stimuli, but the preliminary findings warranted greater support from statistical testing the hypothesis. A more detailed coding scheme developed from the responses from a larger sample was also required to draw more reliable conclusions. Hence, study 2 followed up on the results from study 1, with a major focus on testing differences in interpretations between the two age groups with a larger sample using the six selected stimuli. A detailed coding scheme was also established to code verbatim responses into categories. The method of crowdsourcing was adopted to compare how much a participant’s responses aligned with that of the other participants. Participants’ alignment with their own group and the other group were scored with two separate scoring schemes built on the proportion of group members endorsing each response category.

As it was hypothesised that people tend to align better with similar others than dissimilar others (Apperly et al., 2024), the major prediction of study 2 was that the two age groups would score lower in alignment when being scored with the scoring scheme established with reference to the other age group than when with the scoring scheme established with reference to their own age group.

3.3.1 Method

Participants. Thirty-four younger adults (18-25 years; M_{age} = 21.68) and 34 older adults (53-60 years; M_{age} = 55.97) with balanced gender were recruited via Prolific. The sample size was determined by an a priori power analysis using G* Power (Faul et al., 2009) for achieving .80 power in a paired t-test with a medium effect size (Cohen's *d* = .50) and significance level of .05. Due to the lack of relevant data in the existing literature to determine a likely effect size, an effect size for power analysis was selected based on Cohen's convention and the observation that the median effect size (Cohen's d) reported in meta-analyses of published psychological studies was 0.43 (Lakens, 2022; Richard et al., 2003). The pre-screening criteria were identical to that used in Study 1. None of the new participants had participated in Study 1. Most participants were monolingual (94.12% in the older group; 88.24% in the younger group). In the older adult group, 41.18% of the participants completed a Bachelor's degree or above; in the young group, the percentage was 29.41%. All participants in the older group were White, but ethnicity was more diverse in the younger group: 76.44% of them were White, while the others were Asian (11.76%), Black (5.88%), or Mixed (5.88%). Research Ethics approval was obtained from the Ethical Review Committee at the University of Birmingham. The study was not preregistered.

Stimuli. Six pictures (P3, P4, P5, P8, P9 and P10) were selected from the set of 10 used in Study 1.

Procedure. All questions were administered via Qualtrics. Participants first gave informed consent, then reported their demographics. Instructions stated that they would be shown six pictures in which a person was circled and they had to describe what they thought the circled person was thinking or feeling with complete sentences. As in Study 1, participants could enter multiple interpretations for each picture, but were restricted to entering one interpretation at a time. The six pictures were presented one by one in a randomised sequence. After entering all interpretations for all items, participants were shown the pictures once again, in a randomised order, along with the first five interpretations they entered for each picture. They were asked to rank the interpretations (up to five) according to how likely they thought the interpretation described what the character was thinking or feeling, in descending order (the most likely interpretation ranked first). The total duration of each session was around 30 minutes and participants received £3.75 for completing the study.

Development of coding scheme. A detailed coding scheme was developed using inductive content analysis based on the verbatim responses ranked the most probable by participants.

First, only mental states attributed to the target, but not about the participants themselves or the other character featured in the picture, were extracted as codes. The codes included single mental state terms as well as longer phrases depending on the context, for example "she felt happy about the other person being upset". Multiple codes could be extracted from each response as more than one mental state could be described or implied. For example, "angry" and "worried" were extracted from the response "she is angry and worried". Responses that did not involve mindreading, such as mere descriptions of the character's behaviour, were coded "n/a". After extracting codes from all responses, all unique codes from all six stimuli were listed and grouped into categories based on similarity, forming 25 categories excluding "n/a". For instance, "agitated", "angry", "annoyed" and "fed up" were classified into the "angry/irritated" category. For Study 2, the coding decisions were made solely by the first-coder; inter-rater reliability was evaluated in Study 3. Categories were then also coded by valence, from the point of view of the observer (the participant). For example, “schadenfreude” was grouped in the negative-valence group rather than positive-valence. The complete coding scheme is provided in Appendix A.

Scoring. The scoring schemes were established based on the proportion of participants in each group endorsing each category in their perceived-most-likely responses for each item. This approach of limiting the scoring to the perceived-most-likely responses not only simplified the scoring process, but also ensured a consistent basis for identifying differences in interpretations of the same stimuli. The aim of the scoring procedure was to produce “alignment” scores, which described the agreement of a participant’s response with the pattern of responses in a reference group – in this study, either their own age group or the opposite age group. Comparison of these alignment scores enabled the testing of whether the groups tended to give different interpretations.

The other interpretations entered by participants but not ranked as the most likely description for each item were not considered in the scoring process. To accommodate text responses that involved two or more codes that tapped on distinct categories, each text response was allowed to score on multiple categories. Each response could score a 1 or 0 on up to 25 categories (leading to a vector of 25 binary scores); the n/a category was always excluded in the calculation of alignment scores. Responses that did not involve any mental state interpretation (coded as n/a) were directly given a score of 0.

When the score was calculated based on the proportions of category/valence endorsement in the participant’s own age group, the scoring condition was called “same-group” scoring. Scoring based on the other age group constituted “crossed-group” scoring. A weight was then assigned to each category by calculating the proportion of times it featured in the first-ranked text responses from each age group. The weight was adjusted in the case of same-group scoring to eliminate the problem of data non-independence, by taking away the participant’s data point from the analysis, which will be referred to as the “-1 correction”. In mathematical terms, the weight of each category *j* in the same-group condition was calculated as $\frac{1}{n-\mathbf{1}}\left(\sum_{i=1}^{n} x_{ij} - \mathbf{1}\right)$, where x_{ij} = the binary score (1 or 0) of participant *i* on the category *j* and *n* = total number of participants in the reference group (i.e., the participants' own age group). Then, the binary scores of a response by participant *p* was multiplied with the weight of each category *j* with the formula: x_{pj} $(\frac{1}{n-\mathbf{1}} \cdot (\sum_{i=1}^{n} x_{ij} - \mathbf{1}))$, or equivalently, $\frac{1}{n-\mathbf{1}}(x_{pj} \cdot \sum_{i=1}^{n} x_{ij} - \boldsymbol{x_{pj}})$. By doing so, each response was assigned a weighted-value for each individual category; the weighted-value was either 0 (if the response did not involve the category concerned) or the weight of the category concerned, as calculated in the previous step. A final weighted score was then calculated for each text response. If a response involved only one category, the weighted score was simply the weighted-value of the category that the response involved, whereas for a response that involved two or more categories, this score was the average of non-zero weighted-values across all categories that the text response involved. This weighted score was the participant's same-group alignment score for the item concerned.

The calculation of item scores in the crossed-group condition was similar to the same-group condition, but the calculation was simpler as adjustment for data non-independence was unnecessary in the crossed-group condition. Hence, for each item, the weight for each category *j* was calculated as $\frac{1}{n}\sum_{i=1}^{n} y_{\ ij}$, where y_{ij} = the binary score (1 or 0) on the category *j* of participant *i* from the other age group and *n* = total number of participants in the other age group. The weighted-values for each category assigned to each response by participant *p* was calculated with the formula $\frac{1}{n}\left(x_{pj} \cdot \sum_{i=1}^{n} y_{\ ij}\right)$. The calculation of item weighted score (i.e., crossed-group alignment score for each item) followed the same logic as in the same-group condition.

The weighted scores calculated with each scoring scheme were then averaged across six items to compute two final scores for further analysis, including a crossed-group alignment score and a same-group alignment score, for each participant. A detailed description of the calculation illustrated with a set of toy data is provided in Appendix B.

The calculation of alignment scores was similar at the valence level. Binary scores of categories that were classified in the same valence group were summed up as the valence cell score. Hence, the weighted-value of each response-valence combination that involved mindreading by participant *p* was calculated with the formula: $\frac{1}{k_j n}(v_{pj} \cdot \sum_{i=1}^{n} w_{ij})$ (crossed-group condition) or $\frac{1}{k_j(n-1)}(v_{pj} \cdot \sum_{i=1}^{n} v_{ij} - v_{pj})$ (same-group condition), where v_{ij} = the valence cell score of participant *i* on the valence *j*, w_{ij} = the valence cell score on the valence *j* of participant *i* from the other age group, k_j = number of categories classified in the valence group *j* and *n* = total number of participants in the reference group. The weighted-values that were not 0 for each text response were then averaged to produce a weighted score using each scoring scheme, and the weighted scores on the six stimuli using each scoring scheme of the same participant were then averaged to produce two final alignment scores for final analysis at the valence level.

In summary, the alignment score of a response in each scoring condition depends on the endorsement levels of its coded categories/valences in the reference group. At the category level, a response that includes only a single popular category can sometimes, but not necessarily, score higher than including multiple categories if some of the involved categories in the latter are less endorsed in the reference group. Conversely, tapping on only a single unpopular category could lead to a lower score compared to a response coded on at least one more popular category. Hence, responses that involve at least one popular category and avoid categories endorsed by very few from the reference group would tend to be reap a high alignment score. The same principle applies at the valence level.

3.3.2 Results

Category level. Figure 3.2 shows the number of participants whose first-ranked responses involved each category separately for each age group. By visual inspection, the two age groups showed different patterns of endorsement across categories of mental state interpretations. The bar charts also suggest that most stimuli showed evidence of more than one popular interpretation, sometimes in different valences. The calculation of alignment scores was able to take account of the full distribution patterns, rather than only selecting a single response as the most frequent or "correct" response.

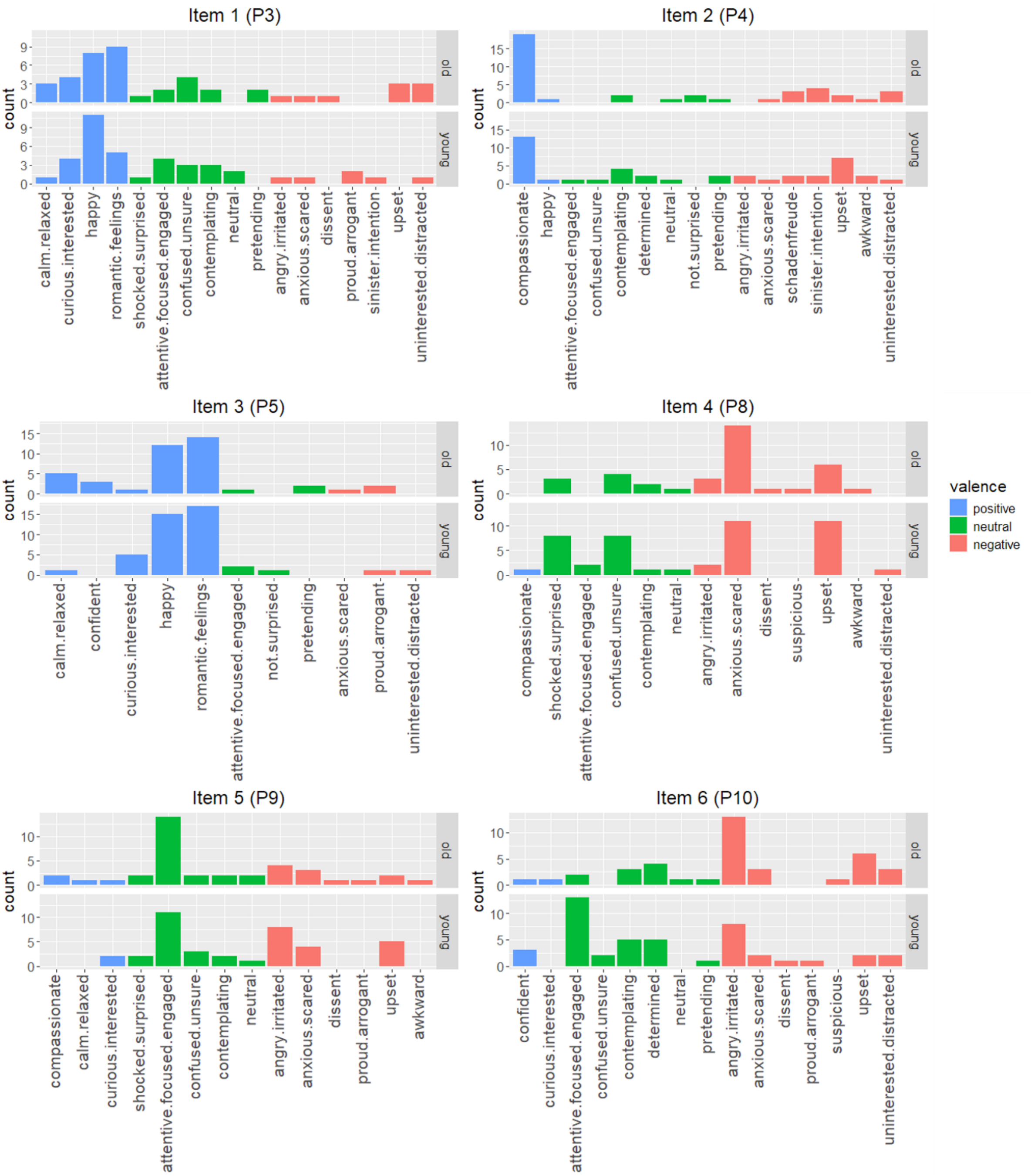


Figure 3.2. Each bar chart shows the number of participants whose first-ranked responses involved each category separately for each age group for the item concerned. The valence of each category is marked by the colour of the bar. Categories that were not endorsed by any participants from either group for each item are omitted from the figure.

Quantitative analysis comparing alignment scores across scoring conditions revealed differences in response patterns between groups by taking into account differences in proportions of endorsement. Table 3.1 summarises the descriptive statistics and paired t-test results of both group in each scoring condition, showing the younger group scoring significantly higher in the young adult (i.e., same-group) condition than in the older adult (i.e., crossed-group) condition. The higher same-group score compared to the crossed-group score in the younger group indicated that younger participants were more aligned with (i.e., had more similar mental state interpretations with) their same-aged peers than with the older participants.

Table 3.1. Summary of descriptive statistics and paired t-test results at category level.

Age group	**Same-group scoring**		**Crossed-group scoring**		**Paired t-test**
	M	***SD***	***M***	***SD***	
Older group	0.208	0.07	0.208	0.06	*t*(33) = -0.04, *p* = .970 Cohen's *d* = -.01
Younger group	0.213	0.04	0.195	0.05	*t*(33) = 2.65, *p* = .012* Cohen's *d* = 0.45

* *p* < .05

Valence level. Paired t-tests at valence level showed that the younger group had significantly higher alignment with other young adults (the same-group scoring) than the older adults (crossed-group scoring), similar to the category level but with a larger effect size. There was no difference within the older group, as shown in Table 3.2.

Table 3.2. Summary of descriptive statistics and t-test results at valence level.

Age group	**Same-group scoring**		**Crossed-group scoring**		**Paired t-test**
	M	***SD***	***M***	***SD***	
Older group	0.079	0.02	0.080	0.02	*t*(33) = -0.01, *p* = .911 Cohen's *d* = -0.19
Younger group	0.084	0.02	0.077	0.02	*t*(33) = 4.00, *p* < .001*** Cohen's *d* = 0.69

*** *p* < .001

3.3.3 Discussion

Study 2 aimed to investigate whether older and younger adults differed in how they interpreted ambiguous social stimuli using differences in alignment scores calculated with reference to two separate age groups. The predicted difference in alignment scores between scoring conditions was observed in the younger group but not in the older group at both category and valence levels. In other words, the younger group aligned more with same-group peers than older adults, but the difference was not observed in the older group. A possible interpretation is that while the younger group were more specifically aligned with same-aged peers, the older age group was more flexible in interpreting mental states in a way that aligned with both people of their age and younger people. However, the way in which alignment scores were calculated leads to an alternative and less informative explanation that cannot be ruled out.

When a category was featured in both groups, the "-1" correction in the same-group scoring condition resulted in a tendency to deflate same-group alignment relative to crossed-group alignment score. This bias was especially strong for participants from the group that had lower endorsement of the categories that were featured in responses from both groups. However, critically, the scoring system could not give rise to false positives (i.e., showing higher same-group alignment than crossed-group alignment as predicted), and could only ever lead to false negatives (i.e., relatively inflating crossed-group alignment and thereby masking the predicted effect) in one of the two groups. Hence, finding a significant difference between scoring conditions in either group, but not necessarily both groups, was sufficient to conclude that the groups differed in interpretations.

To conclude, the current finding supported differences in interpretations of mental states in ambiguous social scenario between the two groups. As the major findings for both category and valence levels did not differ, data were analysed only at the more fine-grained category level in Study 3.

3.4 Study 3: Replication and examining individual differences

To further examine whether there can be multiple legitimate interpretations of ambiguous social stimuli, study 3 was conducted as an attempt to replicate the findings from Study 2 in a larger sample, to explore consistent individual differences in participants' mental state interpretations, and to test whether there were reliable individual differences in alignment and propensity to generate interpretations. More specifically, the distribution of scores was inspected and inter-item correlations were calculated to address this question on individual differences. As higher flexibility could be indicated by being better able to align with dissimilar others in mindreading, alignment scores in the crossed-group scoring condition and the difference in alignment scores between scoring conditions are potential indices of individual differences in mindreading flexibility, but only if stable inter-item correlations are found.

An additional exploratory analysis was conducted by correlating participants' propensity to generate multiple interpretations with their alignment scores, to examine the relationship between propensity to mindread and alignment.

3.4.1 Method

Participants. Eighty-four younger participants and 85 older participants were recruited from Prolific. All participants received £3.75 for completing the study. One participant from each group was screened out because age information was lacking, leading to a final dataset of 83 younger participants (18-26 years, M_{age} = 23.25; 41 female) and 84 older participants (53-60 years, M_{age} = 57.14; 42 female). The sample size was determined by a priori power analysis to detect a correlation of .30 within each group with 80% power at the significance level of .05 using G*Power (Faul et al., 2009). The effect size was chosen based on parallel research in personality suggesting .30 was a threshold for satisfactory inter-item correlation (Epstein & O'Brien, 1985; Mischel, 1968). The same screening criteria as in Studies 1 and 2 were adopted. Additionally, participants who had participated in the previous studies were excluded. Most participants were monolingual (94.0% in the older group; 81.9% in the younger group) with the remaining participants being bilingual (4.8% in the older group and 16.9% in the younger group) or multilingual (1.2% in each group). In the older group, 45.2% of the participants held a Bachelor's degree or above; in the younger group, the proportion was 51.8%. The vast majority of participants in the older group were White (97.6%), while 1.2% was Asian and 1.2% was Black. In the younger group, 68.7%, 21.7%, 7.2% and 2.4% were White, Asian, Black, and Mixed, respectively. Research Ethics approval was obtained from the Ethical Review Committee at the University of Birmingham. The study was not preregistered.

Stimuli. The same six pictures from Study 2 were used in Study 3.

Design and procedures. The design of Study 3 was identical to Study 2; the procedures were almost identical to Study 1. Participants' responses to the 10-item Autism Spectrum Quotient (AQ-10) questionnaire (Allison et al., 2012) were collected but not analysed below. The duration of each session was around 30 minutes and participants received £3.75 for completing the study.

Coding and scoring. The coding and scoring were identical to that in Study 2. Calculation of alignment scores were solely based on participants' first-ranked responses for each item. The only difference was that all responses, not limited to the first-ranked responses, were coded into categories for further individual differences analyses. Inter-rater reliability of coding was evaluated by having a second coder code the first-ranked responses by 34 participants (20% of all participants) with reference to the coding scheme. Inter-rater reliability was satisfactory, with Cohen's kappa ranging from .72 to .83 over all six stimuli.

3.4.2 Results

Comparison to giving random interpretations. Before addressing the question of whether there were multiple common interpretations of the same item, a permutation test was conducted to investigate whether participants showed some extent of alignment in interpreting the same stimulus, or whether interpretations were random such that there was no way to establish at least one interpretation that a significant proportion of participants would agree on. This comparison was made possible by the large sample size recruited in Study 3.

Responses by virtual participants with the sample size of the current full sample (n=167) were simulated by randomly assigning combinations of categories that have been endorsed by at least one participant in the sample across both groups. Item alignment scores were calculated and then averaged across items to form the virtual participant's alignment score. The average alignment scores across all virtual participants were then calculated. This process repeated for 1000 iterations, generating a distribution of 1000 average alignment scores across virtual participants. The 95th percentile of this distribution served as the comparison baseline for the actual average alignment score observed in the sample, pooling both groups together. This comparison baseline was chosen as a nonparametric alternative to parametric tests adopting the .05 significance level. If the observed average alignment score was higher than the 95th percentile of the simulated distribution, it would indicate that participants showed higher agreement than making random interpretations of the items. Observed alignment among participants, *M* = 0.186, *SD* = 0.04, 95% CI = [0.179, 0.192], was significantly higher than the case of making random interpretations, 95th percentile = 0.080 (Figure 3.3).

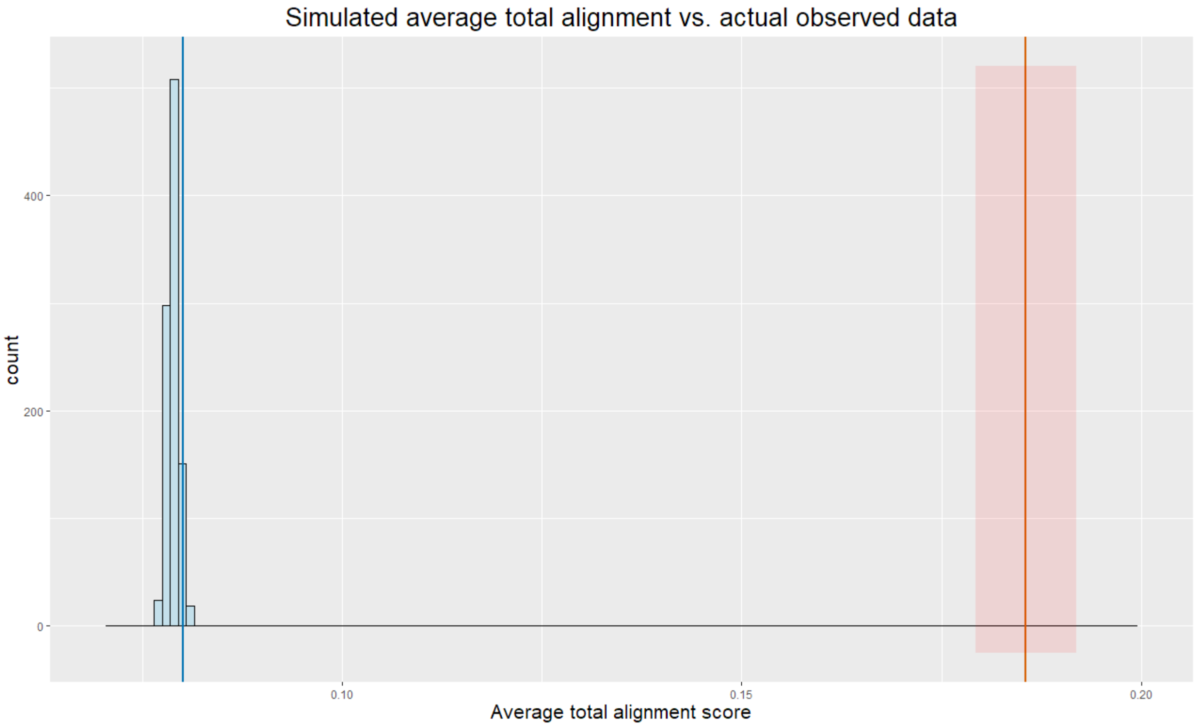


Figure 3.3. The histogram shows the distribution of average alignment in virtual participants simulated over 1000 iterations. The blue line indicates the 95th percentile of the distribution (0.080), which is lower than the red line indicates the observed average alignment across participants in the sample (0.186); the red shaded region indicates the 95% confidence interval of the observed average alignment score estimate.

Permutation tests were also conducted to evaluate whether there were multiple categories that were commonly endorsed for each item. The proportions of virtual participants endorsing each possible category in each of the 1000 simulations were recorded. The 95th percentile of these proportions were compared to the actual proportion of participants (across both groups) endorsing each category. If the actual proportion exceeded the 95th percentile of the corresponding simulated proportion, the category was considered a commonly endorsed category. Results showed that all six items had more than one commonly endorsed category. The number of commonly endorsed categories varied from two to five (Table 3.3).

Table 3.3. List of commonly endorsed categories for all six items in Study 3.

Item	**Category**	**Actual proportion**	**95% percentile of simulated proportion**
P3	Curious/interested	0.169	0.108
	Happy	0.157	0.108
	Romantic feelings	0.145	0.102
	Contemplating	0.211	0.108
P4	Contemplating	0.175	0.108
	Upset	0.151	0.108
P5	Happy	0.434	0.132
	Romantic feelings	0.446	0.132
P8	Shocked/surprised	0.205	0.126
	Contemplating	0.127	0.126
	Angry/irritated	0.133	0.126
	Anxious/scared	0.295	0.126
	Upset	0.193	0.132

P9	Attentive/focused/engaged	0.337	0.102
	Contemplating	0.151	0.102
	Angry/irritated	0.120	0.102
	Anxious/scared	0.163	0.102
	Upset	0.120	0.102
P10	Attentive/focused/engaged	0.223	0.120
	Contemplating	0.277	0.120
	Determined	0.157	0.126
	Angry/irritated	0.331	0.120

Same-group alignment versus crossed-group alignment. Figure 3.4 shows the number of participants whose first-ranked responses involved each category, separately for each age group. Similar to study 2, the two age groups showed some variation in the pattern of category endorsement by visual inspection.

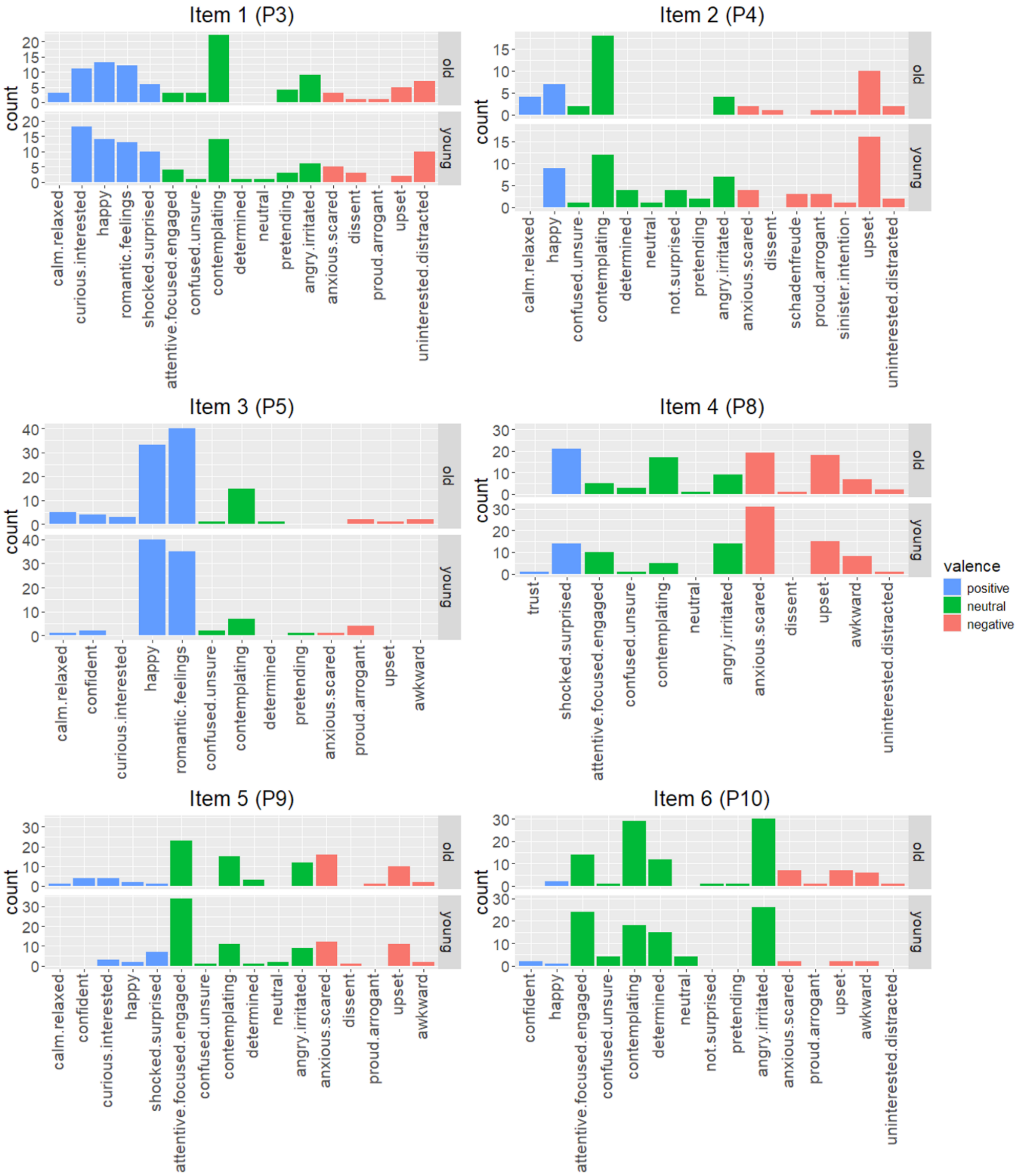


Figure 3.4. Each bar chart shows the number of participants whose first-ranked responses involved each category separately for each age group for the item concerned. The valence of each category is marked by the colour of the bar. Categories that were not endorsed by any participants from either group for each item are omitted from the figure.

Taking into account group differences in proportion of endorsement, the younger group scored significantly higher in the same-group condition than in the crossed-group condition with a small effect, consistent with Study 2, as summarised in Table 3.4. Similar to Study 2, the higher same-group score compared to the crossed-group score in the younger group indicated that younger participants had more similar mental state interpretations with their same-aged peers than with the older participants.

Table 3.4. Summary of alignment scores and paired t-test results between scoring conditions.

Age group	**Same-group scoring**		**Crossed-group scoring**		**Paired t-test**
	M	***SD***	***M***	***SD***	
Older group	0.178	0.04	0.181	0.05	*t*(82) = -0.89, *p* = .376 Cohen's *d* = .10
Younger group	0.200	0.05	0.189	0.04	*t*(82) = 3.01, *p* = .003** Cohen's *d* = .33

** *p* < .01

Individual differences in alignment. The distributions of alignment scores were inspected and compared to a uniform distribution (assuming no variation in alignment score among participants as the null hypothesis) to evaluate whether there was variation in participants' alignment scores. One-sample Kolmogorov-Smirnov tests showed that same-group and crossed-group alignment scores in both groups deviated from a uniform distribution, *D* = .17 to .30, *p* varied from <.001 to .016. The distributions of observed scores are shown in Figure 3.5.

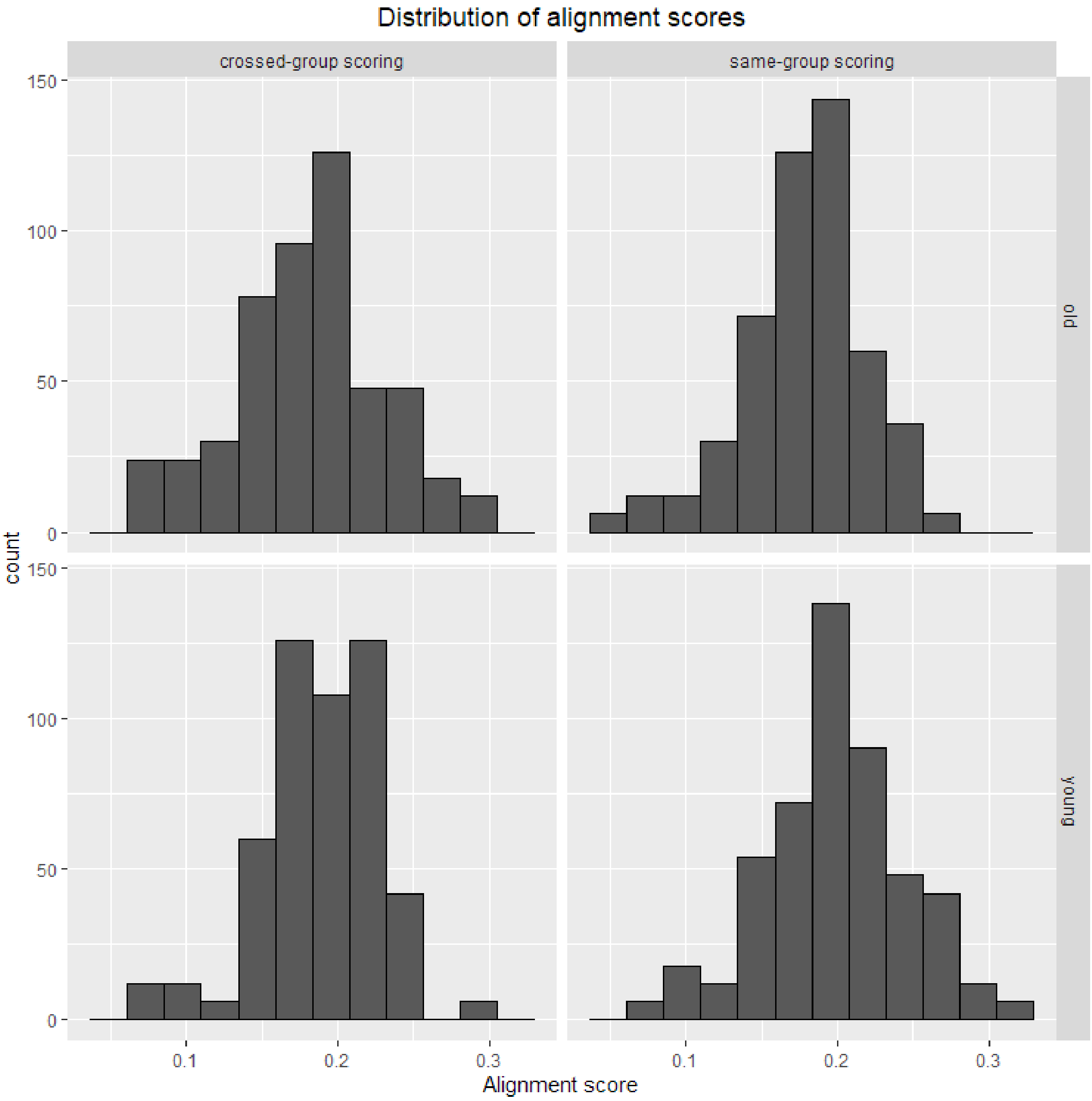


Figure 3.5. Variation was observed in the alignment scores in both scoring conditions in both groups.

The inter-item correlations in alignment scores were then inspected to evaluate whether participants had consistent levels of same-group and crossed-group alignment across items, as well as difference scores calculated by subtracting same-group alignment from crossed-group alignment scores. The difference scores could indicate a participant's alignment with the crossed-group controlling for their alignment with the same-group. The zero-order Pearson correlations are summarised in Tables 3.5 to 3.7. With reference to the criterion of using .30 as a threshold for satisfactory inter-item correlation from parallel research in personality (Epstein & O'Brien, 1985; Mischel, 1968), the inter-item correlation of alignment scores were not satisfactory as none of the correlations exceeded the threshold, except the correlation between difference scores in P3 and P4 (*r* = .32) in the younger group. In other words, greater alignment with same-age or other-age adults on one item was not reliably correlated with greater alignment on other items.

Table 3.5. Zero-order Pearson correlations among same-group (below diagonal) and crossed-group (above diagonal) alignment scores in the older group.

	P3	**P4**	**P5**	**P8**	**P9**	**P10**
P3	—	-.23*	-.13	.02	.01	-.28**
P4	.03	—	.06	.19	-.03	.12
P5	.07	.11	—	.13	.11	-.05
P8	-.04	.07	-.06	—	.18	.00
P9	.09	-.08	-.02	.15	—	-.06
P10	-.12	.12	-.09	.29**	.00	—

* *p* < .05

** *p* < .01

Table 3.6. Zero-order Pearson correlations among same-group (below diagonal) and crossed-group (above diagonal) alignment scores in the younger group.

	P3	**P4**	**P5**	**P8**	**P9**	**P10**
P3	–	.15	.17	.07	.05	-.15
P4	-.06	–	-.08	.00	.03	-.02
P5	-.01	.05	–	.09	-.04	.04
P8	.10	.02	.08	–	.01	-.01
P9	-.04	.05	-.07	.00	–	-.17
P10	-.17	.12	.08	-.01	-.10	–

Table 3.7. Zero-order Pearson correlations among difference scores in the younger group (below diagonal) and the older group (above diagonal).

	P3	**P4**	**P5**	**P8**	**P9**	**P10**
P3	–	.07	-.27	.10	-.09	.00
P4	.32**	–	.09	.07	-.05	.09
P5	.20	.04	–	.09	.10	.25
P8	.10	-.04	-.01	–	.13	.21
P9	.05	.06	.12	.11	–	.09
P10	.03	.22	..06	.11	.27	–

** *p* < .01

Individual differences in propensity to generate multiple interpretations. The number of unique response category combinations for each stimulus was averaged across the six items. The average number of unique categories varied from 1 to 3.33 in the younger group, *M* = 1.87, *SD* = 0.60. That of the older group varied from 1 to 4, *M* = 1.91, *SD* = 0.73.

One-sample Kolmogorov-Smirnov tests showed that the distributions in both groups significantly deviated from a uniform distribution, D = .26 to .37, p<.001, also indicated in Figure 3.6.

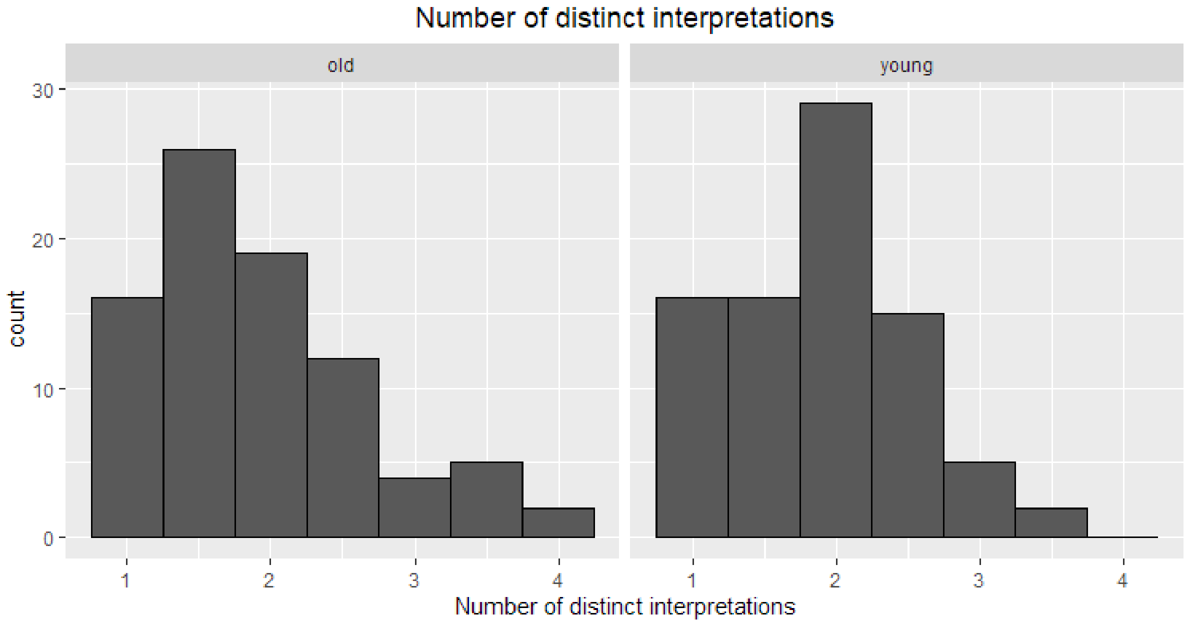


Figure 3.6. Variation among participants was observed in their average number of distinct interpretations (number of unique category combinations) across items.

Inter-item correlations were satisfactory, with Spearman correlations exceeding .30 except for two pairs of items (P5 and P9, r_s = .28; P9 and P10, r_s = .29); all correlations were significant. The pattern and statistical significance of inter-item correlations was not altered when controlled for average word count in participants' responses across items. The inter-item correlations are summarised in Table 3.6. Cronbach's alpha was .83, indicating good internal consistency.

Table 3.8. Zero-order spearman correlations (controlled for verbosity) between items.

	P3	**P4**	**P5**	**P8**	**P9**
P4	.55*** (.55)				
P5	.46*** (.45)	.41*** (.41)			
P8	.50*** (.50)	.47*** (.47)	.53*** (.52)		
P9	.50*** (.49)	.38*** (.37)	.28*** (.26)	.39*** (.38)	
P10	.47*** (.47)	.42*** (.42)	.44*** (.44)	.43*** (.43)	.29*** (.28)

*** *p* < .001

Correlation between alignment and propensity. Each participant's average number of unique interpretations generated across items was calculated as an overall index of mindreading propensity. Their average same-group alignment scores, average crossed-group alignment scores, and average alignment with the whole sample pooling both age groups across items were also calculated. The correlations between propensity and average alignment with the same-group, *r* = -.07, *p* = .393, and the pooled sample, *r* = -.12, *p* = .137, were non-significant. A weak negative correlation between propensity and crossed-group alignment was found, *r* = -.15, *p* =.049. However, this correlation was not statistically significant when Bonferroni correction was applied to account for the three exploratory correlation analyses.

3.4.3 Discussion

Study 3 aimed to investigate whether having multiple popular interpretations of social ambiguous social stimuli was common, to replicate the differences in mental state interpretations by older and younger adults, and to explore possible indices of individual differences in mental state interpretations. Results showed that more than one popular interpretation was observed across all stimuli. Similar to Study 2, the younger group had significantly higher alignment with their own age group than the older age group, which supported the idea that different groups of individuals may interpret the same social stimuli in different ways, analogous to findings showing differences in mental state interpretations between groups differing in culture (Adams et al., 2010; Perez-Zapata et al., 2016) and neurotype (Alkhaldi et al., 2019; Edey et al., 2016; Milton, 2012). Number of distinct interpretations, but not alignment scores, was consistent across stimuli, rendering it a possible candidate of indicating individual differences in mindreading.

Comparison to giving random interpretations. Permutation test results indicated that participants agreed on certain mental state interpretations, but there were multiple popular interpretations for each item. These findings challenge the assumption that the "correct" answer decided by experimenters or pilot participants serves as a proxy for a single best interpretation that is invariant across different groups of individuals. This is especially relevant to tasks requiring participants to interpret mental states in ambiguous social scenarios as the actual thoughts or feelings of the targets are unknown.

Same-group alignment versus crossed-group alignment. The younger group showed higher alignment with their own group than with the older group, corroborating the major finding from Study 2 that interpretations differed between the two groups. Furthermore, the effect was not solely driven by differences in the endorsement of the most popular category between the two groups, but also by differences in the endorsement of other categories, as shown in Figure 3.4. As the permutation test result reported above has revealed the existence of multiple popular categories, taking into consideration the endorsement of categories other than the most popular one provides a more comprehensive evaluation of how the two groups interpreted the same stimuli in different ways.

Investigating stable individual differences. Alignment scores and propensity to generate multiple unique interpretations were investigated as potential facets of consistent individual differences in mindreading. Both variables showed significant variation between individuals. However, reliable individual difference was only found in the number of distinct interpretations participant suggested across items, but not their alignment score with either same-group or crossed-group scoring. These results suggest that the former but not the latter is a potential candidate in indicating individual differences in mindreading.

Correlation between alignment and propensity. The exploratory examination of the relationship between alignment and propensity revealed that they were likely orthogonal. In other words, having a higher tendency to generate multiple mental state interpretations did not imply being better at agreeing with other members of a group. This finding corroborates the wider literature suggesting propensity and accuracy of mindreading are independent constructs (e.g., Carpenter et al., 2016), although alignment is considered a way of considering appropriateness of a mindreading response alternative to accuracy in the current study.

3.5 General discussion

3.5.1 Summary

The overarching aim of this chapter was to demonstrate the presence of multiple interpretations of ambiguous social stimuli, and examine individual differences in the generation of these interpretations. Multiple popular interpretations were observed for each stimulus and older and younger adults showed different patterns of interpretations indicated by alignment scores. Although the alignment score was not a reliable index of individual differences in mindreading, the number of distinct interpretations was a possible candidate as it demonstrated satisfactory inter-item correlation.

All Studies 1 to 3 demonstrated notable variation in how people interpret mental states in ambiguous social scenarios, even within neurotypical adults. In Study 3, simulation results further showed that there were multiple popular interpretations for every item. These challenge the assumption in existing tasks that a single correct answer decided by the experimenters or the pilot panel members is representative of a consensus among neurotypical individuals, which then serves as a proxy of ground truth. The calculation of alignment scores has provided a novel way to quantify participants' open-ended response patterns, which can be adopted to study interpretations by individuals who differ in other aspects, for example, gender, ethnicity, and clinical conditions. It can be potentially utilised to study the double empathy problem between neurotypical individuals and the autistic population as well.

Despite identifying differences in interpretations across age groups, alignment scores or differences in alignment between scoring conditions were not reliable indices of individual differences due to low inter-item correlations. The reason was unclear. It was possibly because aligning with other people in making mental state attribution in ambiguous social scenarios was not a stable trait but instead depended on the stimulus itself. It was also possibly due to the current open-ended format of the task as well as the ambiguity of the scenarios with no background information provided, which permitted a broad range of interpretations, making it difficult for participants to always align well with others across scenarios. A forced-choice format, where participants select the most plausible interpretation from a given set rather than generating multiple ones on their own might enhance consistency in alignment across items. Similarly, when more background information about the interaction in the scenarios is provided, participants may show more consistent alignment across scenarios. These questions will be addressed in Chapter 5.

The exploration of participants' propensity to offer multiple unique interpretations as a factor of individual variance in mindreading revealed consistency across items. However, this facet of mindreading was more similar to propensity to mindreading or motivation than flexibility of adjusting interpretations with context, the focus of the current thesis. It is recommended as a point for future research but not for the current thesis.

3.5.2 Limitation

A notable limitation of the current set of studies was that the split by age was arbitrary. Participants can be split into groups in many other ways, for example, by gender or ethnicity. However, any reasonable despite arbitrary grouping is informative as any differences found between two groups of neurotypical adults would challenge the assumption of a single ground truth in mental state interpretations across all neurotypical individuals.

Another limitation is that although the current results showed multiple popular interpretations for each item, these findings did not rule out the possibility that one interpretation, or some interpretations, were more “accurate” than others, despite the true mental states of the target were not directly accessible. This possibility is further addressed in Chapter 5 in which context is manipulated, providing support for the view that the alternative popular categories are likely alternative legitimate ways of interpreting the stimuli, rather than fallacies.

3.5.3 Conclusion

This chapter found that ambiguous social stimuli elicited multiple popular interpretations, with older and younger adults showing different patterns. Alignment scores were not reliable for indicating individual differences, but the number of distinct interpretations was promising, showing good inter-item correlation. The results challenge the assumption of a single correct answer in mindreading tasks.

Chapter 4

Generation and recognition in mental state interpretation

4.1. Introduction

In Chapter 3, Studies 1 to 3 featured an open-ended task that required participants to provide mental state interpretations of pictorial stimuli and found differences in such interpretations between older and younger adults. However, even for the same social or demographic group, altering the task format may affect how participants interpret the ambiguous social stimuli, because the processes required to recognise a good interpretation is likely different from generating a good interpretation for oneself. This chapter aims to compare participants' preferences for mental state interpretations between task formats.

Research in social cognition has revealed that typically developing children and children with learning disorders performed better in the conventional forced-choice version of the Reading the Mind in the Eyes Task (RMET) compared with an open-ended version (Cassels & Birch, 2014). The authors attributed the result to that the forced-choice format allowed for compensatory strategies, for example, by elimination, to arrive at the correct answer. This finding parallels research on memory, which has a long history of showing that participants' performance in recognition tasks often differed from free-recall tasks, and memory deficits were less likely to be detected with recognition tasks (e.g., Breen, 1993; Calev, 1984). Specifically, in the studies by Tulving and Walkins (1973), the authors found that participants were more likely to recall a list of five-letter words with an increasing number of memory cues provided, or to recognise whether a word was present in the list, compared to free recall. Such results suggested that recognition facilitated memory retrieval by providing cues, unlike free recall tasks where no cues were available. Some other studies suggested that recognition and free recall engaged differential cognitive processes as individuals with Parkinson's disease (Breen, 1993) and schizophrenia (1984) only showed memory deficits when tested on recall but not recognition. These studies provide insight into how participants could possibly perform differently in highly similar tasks that vary in task format.

For the current task, the forced-choice format resembles recognition tasks, while the open-ended format is akin to free-recall tasks. Although participants did not recall information being presented, unlike memory tasks, varying the response format changed the cues available for interpreting the same social scenario. In the forced-choice format, participants chose from a limited set of choices, which could act as cues, unlike the open-ended format where participants had to generate interpretations without cues. Individuals also showed higher propensity to attribute mental states when presented with the forced-choice version of the RMET compared to the open-ended version of it (Betz et al., 2019). Hence, the difference in task format, though seemingly trivial, could influence participants' tendency to generate mental state interpretations as well as their decision on which interpretation of the social scenario was perceived to be the most plausible.

Comparing participants' interpretations of the same stimuli across different task formats also addresses an important gap in the existing mindreading literature. Chapter 2 highlighted the inconsistencies in administering the same tasks with different formats across studies. For example, the Animations task (Abell et al., 2000) was adopted in a forced-choice format in some studies (e.g., Brewer et al., 2017; 2022) and in an open-ended format in others (e.g., Kéri et al., 2020; Livingston et al., 2021). Some researchers argue that open-ended formats are more sensitive to individual differences, are more naturalistic and are more likely to capture perspective-taking-specific processes (e.g. Cassels & Birch, 2014). However, the forced-choice format is popular for its convenience and ease of implementation. Typically, forced-choice tasks in the existing literature assume a model answer with other alternatives serving as foils. This design has been criticised for potentially allowing participants to select the model answer by eliminating unlikely foils, rather than actively interpreting the target's mental states (Cassels & Birch, 2014). This argument is corroborated by research showing that clinical groups with social deficits or groups high in psychopathic tendencies related to social deficits recruited other cognitive strategies to solve mindreading tasks with non-mentalistic strategies (Gordon et al., 2004). Furthermore, it should not taken for granted that the model answer, usually decided upon by the experimenter or piloted with a small group of participants or experts, remains the "best description" endorsed by most people when task format is altered. This method for generating the model answer assumes the decision by the researchers, pilot participants, or experts, is generalisable to the population regardless of group membership or task format. However, if this assumption does not hold true, current practices become problematic, which necessitates caution in comparing or aggregating task performance across different formats of notionally the same tasks. Therefore, it is crucial to scrutinise whether two assumptions are valid: that (1) forced-choice mindreading tasks do require participants to engage in mindreading and (2) the model answers remain consistent across different tasks formats.

With the current study, the impact of task format was examined by presenting alternative interpretations derived from actual verbatim responses collected from the open-ended Study 3, which represented plausible and likely interpretations alongside the most popular ones. In other words, the alternatives in the current task were not arbitrary foils, but genuine possible alternative interpretations of the stimuli. This approach, thus, reduced the likelihood of participants completing the task without actively engaging in mindreading but adopting the non-mentalistic strategy of elimination, because the alternative were unlikely to be easily eliminated. This approach also allowed for assessing if the proportion of participants endorsing each of the four alternative interpretations was influenced, and whether the most popular interpretation was changed to another plausible interpretation, when a forced-choice format was administered in contrast to the original, open-ended format. If such changes in responses were observed, the results would challenge the practice of using the same scoring scheme without considering if the answer should change when the task was adapted to another format.

The current study aims to address the major question of whether it is valid to assume a single typical way to interpret the mental states of a target in ambiguous social stimuli among neurotypical adults. If such a consensus exists, the most popular interpretation of each stimulus should remain consistent, regardless of changes in response formats. This would be indicated by similar patterns of endorsement of mental state interpretations across implementing the task in either forced-choice or open-ended formats. Alternatively, if the perceived-most-plausible interpretation changes with the response format of the task, participants' decision on the best interpretations would be expected to align more with other participants who took the task presented in the same format, in comparison to another format. Specifically, the endorsement of various interpretations from participants who responded to a forced-choice format of the task (in the current sample) were compared with that of participants who responded to the open-ended version of the task (the young group sample in Study 3), by calculating current participants' alignment scores based on the two separate reference samples. In other words, two sets of scores for the same (current) sample were calculated and compared. It was predicted that test format would influence participants' interpretation of the stimuli.

Additionally, if certain interpretations of the same stimulus are commonly perceived as better than other plausible interpretations, the endorsement of various plausible interpretations should diverge from a chance distribution, and at least one plausible interpretation should be endorsed by a significantly higher proportion of participants than zero.

4.2. Method

The current study was pre-registered on OSF prior to data collection (https://osf.io/4kh2g).

4.2.1 Participants

Forty-four participants aged from 18 to 25 (22 female, M_{age} = 22.41) were recruited for the current study, which was the required sample size to detect an effect with a hypothesised effect size of w = 0.5 with a χ^2 goodness of fit test with .80 power at $\alpha = .05$ as indicated by G*Power (Faul et al., 2009). This effect size was chosen because only a substantial deviation from a chance distribution in participants' interpretation preferences was meaningful to serve as a minimum baseline to set up subsequent analyses on any shifts in preferences when the task format was manipulated. All participants were recruited online via Prolific. The following screening criteria were applied: participants had to be UK residents, spoke English as their first language, had not been diagnosed with ASD, and had not participated in the previous series of studies. Among the participants, 90.9% were monolingual, 6.8% were bilingual and 2.3% spoke more than two languages. Around half of the participants had not obtained a Bachelor's degree (47.8%). Participants identified their ethnicity according to the descriptions recommended by the United Kingdom Office for National Statistics (ONS). Most of the participants were White (86.4%); 6.8% were Asian, 4.6% were mixed, and 2.3% were Black.

The demographics of the current sample and the younger group sample from Study 3 are summarised below in Table 4.1.

Table 4.1. Summary of demographics of the participants from the younger group sample from Study 3 and the current sample (Study 4).

Demographics	**Younger group from Study 3 (n=83)**	**Current sample (Study 4) (n=44)**
Age	*M* = 23.25 (Range = 18 – 26)	*M* = 22.41 (Range = 18 – 25)
Female proportion	49.4% (n = 41)	50% (n = 22)
Monolingual proportion	81.9% (n = 68)	90.9% (n = 40)
Proportion holding a Bachelor's degree	39.8% (n = 33)	52.2% (n = 23)
Proportion of White individuals	68.7% (n = 57)	86.4% (n = 38)

4.2.2 Study design and procedure

Informed written consent, approved by the Ethical Review Committee at the University of Birmingham, was obtained online before all participants participated in this study. They completed an online questionnaire on the Qualtrics survey platform, in which they were shown six pictures, each depicting a naturalistic social scenario. Each page only contained one picture, the instruction asking participants to select the alternative that they thought best described what the target character (circled) was thinking or feeling, and the four alternatives. Participants were only allowed to choose one option for each question, and they had to submit their response before they could proceed to the next picture. The presentation order of the stimuli was randomised.

Participants were not provided any information about the context of the social situations presented. They were required to select one of four options given, each describing one possible interpretation that they thought best described what the target character was thinking or feeling. The four options for each stimulus were presented in random order, varying between participants. Participants also rated their confidence in their chosen option to be best describing what the target character was thinking or feeling in comparison to the other three options on a 7-point likert scale varying from 1 (not certain at all) to 7 (absolutely certain).

Testing was completed in one session and the duration of the session was around five minutes. Participants received £1.5 after completing the study.

4.2.3 Materials

The same six pictures from Study 2 and Study 3 were used in this study. These pictures depicted various ambiguous social scenarios. The four alternative options were derived from actual verbatim responses given by participants in Study 3, reflecting diverse interpretations of the depicted scenarios.

The primary principle of selecting alternatives for the current forced-choice task was to select the top four popular categories from the open-ended data. As detailed in the previous chapter, participants' verbatim responses were coded into categories using an established coding scheme, and each response could involve multiple categories. Among responses involving multiple categories, there were unique combinations of coded categories (category-combinations). For example, "romantic feelings and happy" was considered a category-combination distinct from "romantic feelings and curious/interested". When selecting alternative options for the forced-choice task, only responses coded with a single category (e.g., "romantic feelings") were considered. This approach affected the selection of alternatives for two items, P4 and P5: the popular category-combinations "compassionate and contemplating" was excluded for P4, while "romantic feelings and contemplating" and "romantic feelings and happy" were excluded for P5. As these excluded combinations involved at least one category (“compassionate” for P4; “romantic feelings” and “happy” for P5) already included in the list of top-four popular categories, this method did not introduce bias in selecting alternatives for the current forced-choice task. One representative verbatim response for each distinct chosen category was then chosen to be presented as the corresponding alternative option in the forced-choice task (e.g., “The man is in love with his partner” for the category “romantic feelings”). The responses were slightly modified, primarily by expanding brief answers into complete sentences, to ensure that all alternatives were similar in presentation style.

4.3. Results

4.3.1 Participants’ confidence in their choices

Participants rated their confidence in their choices for each item on a likert scale from 1 to 7 and their ratings. The descriptives are summarised in Table 4.2. Averaging across items, participants were moderately confident (*M* = 4.39, *SD* = 0.95) that their chosen options best described what the target was thinking or feeling in comparison to the other three options.

Table 4.2. Participants' ratings of their confidence in their chosen options best described what the target was thinking or feeling on a 7-point likert scale (1 = not certain at all, 7 = absolutely certain).

Item	**Range**	**Mean**	***SD***
P3	1 - 7	4.11	1.43
P4	1 - 7	4.41	1.65
P5	2 - 7	5.05	1.08
P8	2 - 7	4.2	1.41
P9	1 - 7	4.18	1.48
P10	1 - 7	4.41	1.39
Average across items	2.33 - 6.17	4.39	0.95

4.3.2 Changes in rank-order preferences of interpretations between response formats

Preliminary inspection of interpretation preferences. To examine if participants perceived that some interpretations were better than the others for each item, the distribution of participants' endorsement of each of the four options in the current sample was compared to a chance distribution with a series of χ^2 goodness of fit tests. Results showed that the distributions for all six items significantly deviated from a uniform distribution (for P3, $\chi^2(3) = 10.36$, $p = .016$; for the remaining five items, $\chi^2(3)$ ranged from 17.64 to 36.55, all *p*s < .001). Subsequent exact binomial tests (Table 4.3) showed that the proportion of participants endorsing the most popular option for all six items significantly deviated from zero (all *p*s < .01), providing further evidence that participants showed agreement in which interpretation was better than the others.

Table 4.3. Number of participants selecting each option as the most plausible interpretation and Binomial exact test results.

Item	**Category represented by option**	**Observed frequency**	***p***
P3	Curious/interested	19	.006*
	Romantic feelings	4	.998
	Contemplating	10	.691
	Happy	11	.558
P4	Compassionate	13	.294
	Upset	2	.999
	Happy	23	<.001*
	Angry/irritated	6	.979
P5	Happy	12	.420
	Romantic feelings	27	<.001*
	Contemplating	2	.999
	Calm/relaxed	3	.999
P8	Anxious/scared	16	.063
	Shocked/surprised	20	.003*
	Upset	4	.998
	Angry/irritated	4	.998
P9	Attentive/focused/engaged	23	<.001*
	Anxious/scared	6	.979
	Angry/irritated	7	.948
	Upset	8	.892
P10	Angry/irritated	8	.892
	Attentive/focused/engaged	12	.420

Contemplating	2	.999
Determined	22	<.001*

* Significant after Bonferroni correction.

The endorsement of categories corresponding to the four options for each item is depicted in Figure 4.1. Qualitative inspection of the rankings of the four categories for each item based on proportion of endorsement showed that the highest-ranked category changed for four items, P4, P5, P8 and P10. The calculation of proportions of endorsement for the open-ended sample is explained in the subsequent subsection on alignment score analysis.

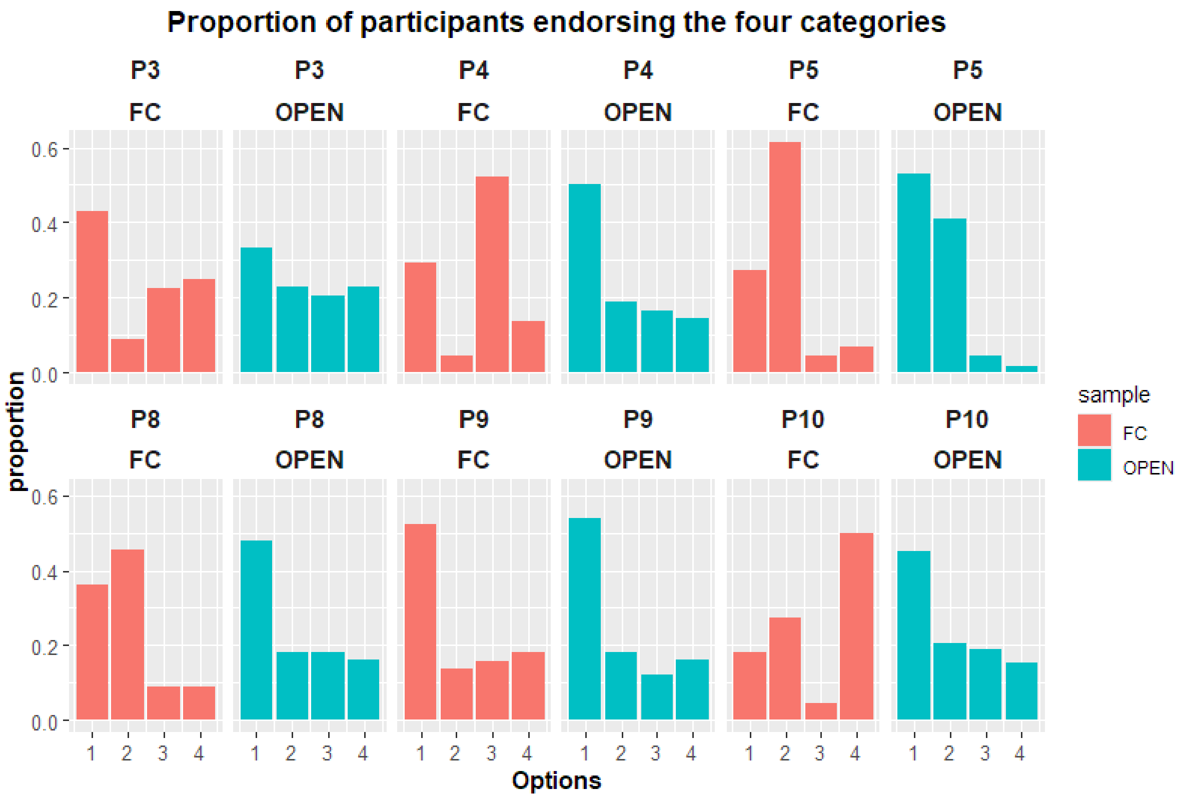


Figure 4.1. The proportion of participants from the current sample (FC, standing for forced-choice) and from the sample of Study 3 (OPEN, standing for open-ended) are presented side-by-side. The patterns of endorsement differed across the two samples.

These results established that participants preferred particular interpretations to other options in the forced-choice task, and the next step was to examine whether their preference had significant differences with those who took the open-ended task.

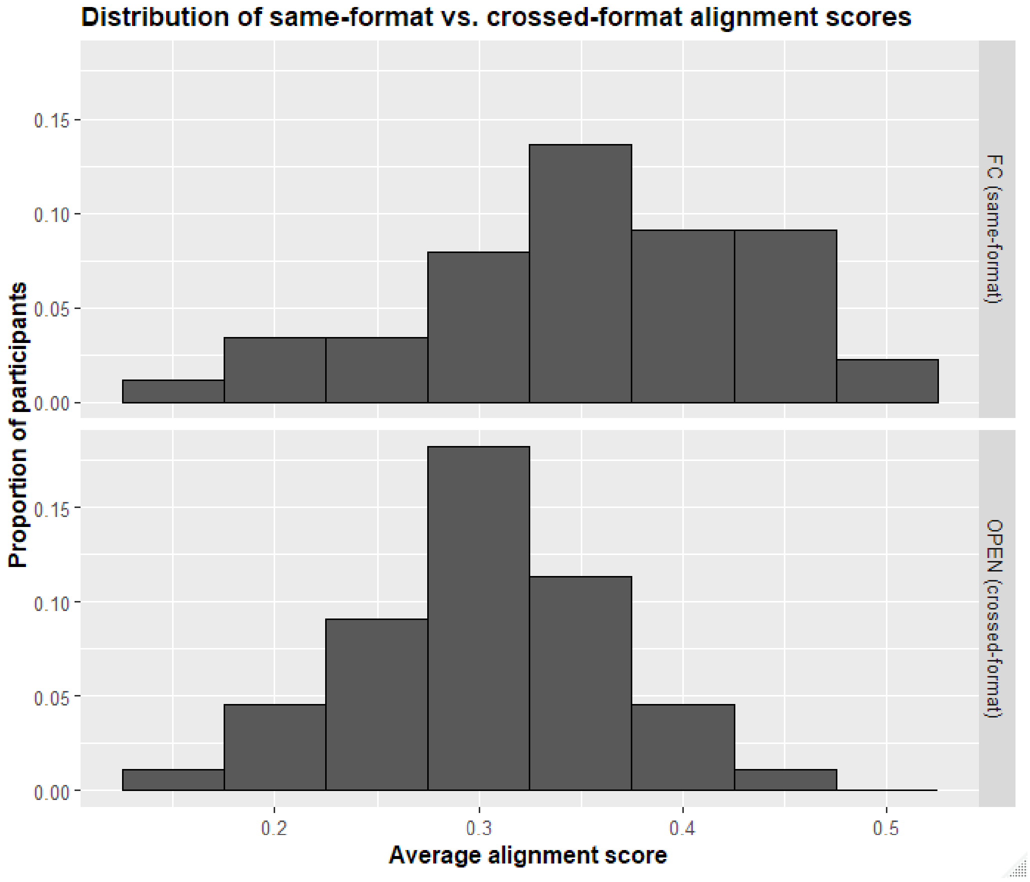
Alignment score analysis. To examine whether there was always a single typical way to interpret ambiguous social stimuli across task formats, participants' agreement with those who responded in a forced-choice format (i.e., the current sample; same-format sample) were compared with their agreement with those who took the task in an open-ended format (i.e., the younger adult sample from Study 3; crossed-format sample).

In the current study, participants only selected one option, corresponding to one category, out of four alternatives for each item. Participants' same-format and crossed-format alignment scores for each item was the proportion of other participants in the corresponding reference sample who endorsed the same category. Hence, each participant's response to an item received two alignment scores: one same-format and one crossed-format. Same-format alignment scores across six items were averaged to produce an overall same-format alignment score for each participant, and the same principle applied for calculating the overall crossed-format alignment score for the participant. If the pattern of category endorsement did not differ between the two samples, the resulting average same-sample and crossed-sample alignment scores should not significantly differ.

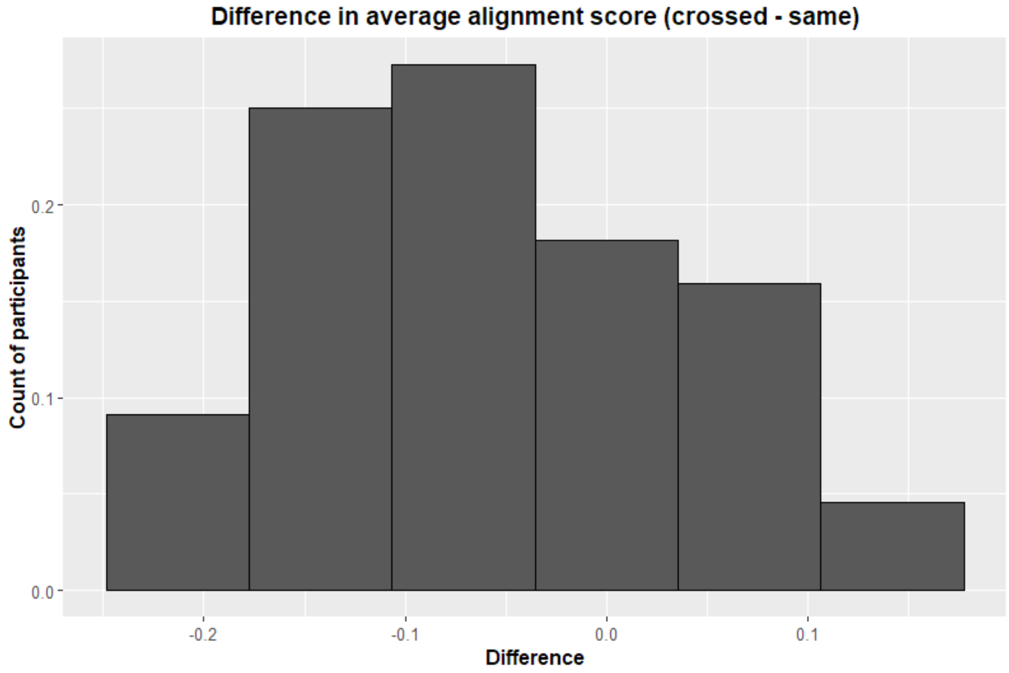
The calculation of alignment scores are explained in more detail below. For each participant, the same-format alignment for each item was calculated with the formula $\frac{number\ of\ participants\ in\ the\ current\ sample\ choosing\ the\ same\ option-1}{total\ number\ of\ participants\ in\ the\ current\ sample\ -1}$ (the -1 adjustment was conducted to eliminate the problem of non-independence of data); the values were then averaged across all six items as the participant's average same-format alignment score.

The crossed-sample alignment for each item was the proportion of participants in the open-ended sample whose first-ranked response involved solely the category corresponding to the option chosen by the current participant. The denominator for calculating such a proportion was determined by summing the number of participants in the open-ended sample who endorsed the four categories corresponding to the four alternatives for the item in the current forced-choice task. As explained in the Materials subsection, among the open-ended responses, only those coded on a single category, but not category-combinations, were considered. For example, if participants were presented options corresponding to categories A, B, C, and D for an item in the current study, and a participant chose the option corresponding to category A, this participant's crossed-group alignment score for the item was calculated as $\frac{number\ of\ participants\ in\ the\ open-ended\ sample\ endorsing\ only\ A}{total\ number\ of\ participants\ in\ the\ open-ended\ sample\ endorsing\ only\ A,B,C\ or\ D}$. These scores were then averaged across the six items for each participant to obtain each participant's average crossed-format alignment score.

The distributions of the two average alignment scores are shown in Figure 4.2(a) and the distribution of the difference between the two average alignment scores is shown in Figure 4.2(b). A paired-sample t-test was conducted to compare participants' average same-format and crossed-format alignment scores. Result showed that participants' average crossed-format alignment scores (*M* = 0.30, *SD* = 0.06) were significantly lower than their average same-format alignment scores (*M* = 0.35, *SD* = 0.08) with a moderate effect size, *t*(43) = -4.01, *p* < .001, Cohen's *d* = 0.61.



(a)



(b)

Figure 4.2. The distribution of the (a) two average alignment scores and (b) the distribution of their within-participant differences are approximately normal.

4.4. Discussion

The primary aim of the current chapter was to examine if changing the task format from open-ended, which required generation of mental state interpretations, to forced-choice, which involved selection but not generation, would influence participants' preferred mental state interpretations of the same social stimuli. Results showed that participants had a preference for specific mental state interpretations for all items and the most popular option was different between the two task formats for four out of six items. Furthermore, participants' overall response patterns aligned better with those who also completed the forced-choice task than those who completed the open-ended task.

4.4.1 Summary

Preference for interpretations. The result showing participants' consistent above-chance preference for particular interpretations suggests that even when having to choose between competitors that were considered likely interpretations, participants still tended to favour certain interpretations over others. However, it was a crucial finding that participants' preferences were influenced by task format, as the most popular interpretation shifted in four out of six items when task format was changed from open-ended to forced-choice. The effect of task format was addressed more rigorously with the alignment score analysis.

Better alignment with forced-choice sample. The result showing higher alignment with the forced-choice sample than the open-ended sample provides support for the hypothesis that task format influences one's interpretation of ambiguous social stimuli. These results parallel findings from memory research showing better performance in recognition versus free recall tasks in both neurotypical and neurodivergent populations (e.g., Breen, 1993; Calev, 1984; Tulving & Walkins, 1973), as well as relevant social cognition research showing better performance in the forced-choice version compared to the open-ended version of task (Cassels & Birch, 2014). However, it is important to note that the current study fundamentally differed from these previous studies. Instead of testing differences in performance levels ("accuracy") between responding to the two formats, the current study focussed on participants' preferences for mental state interpretations.

The divergence in focus can be illustrated by comparing the current study with Cassels and Birch's (2014) study, which adapted the originally forced-choice RMET into an open-ended format and scored a response based on whether its valence matched the valence of the correct answer in the forced-choice task. Hence, their open-ended task allowed for a wider range of "correct" answers, while keeping the distinction between "correct" from "incorrect" responses. Their study focused on comparing task difficulty, indicated by participants' overall performance on hitting the "correct" answers. In contrast, the current forced-choice task was derived from an open-ended task, and the options were legitimate interpretations favoured by participants who completed the open-ended task. There was no strict distinction between "correct" and "incorrect" answers. Instead, the current study indicated that a presumed invariant "correct" answer may not hold true when the task format changed if "accuracy" was based on consensus. This contrasts sharply with the assumption made in the study by Cassels and Birch (2014).

4.4.2 Limitations

A concern with the current task's switch from open-ended to forced-choice format is that it potentially reduced sensitivity in detecting participants' true mindreading ability, here operationalised as alignment with others in one's interpretation of the targets' mental states. This is because there is higher likelihood that participants engage in non-mentalistic strategies such as elimination in forced-choice settings. This criticism, however, was less pertinent to the current task than to typical forced-choice tasks, as the alternatives in the current study were genuine, plausible interpretations generated by previous participants, rather than arbitrary foils. The construct captured by the current task was the extent to which participants' interpretation aligned with that of the others, rather than identifying a single “correct” answer among multiple incorrect options. Furthermore, the forced-choice format is advantageous for designing future studies focused on examining participants’ flexibility in mindreading as a source of individual differences, as will be detailed in the subsequent chapter.

4.4.3 Conclusion

In conclusion, the current study suggests potential issues in adhering to a consistent “correct” mental state interpretation of ambiguous social stimuli across different task formats. Results indicate that there is often a “typical” interpretation perceived as the best descriptor of a target’s thoughts or feelings in a given social situation, although the interpretation is influenced by test format.

Chapter 5

Effect of context on selection of mental state interpretation

5.1. Introduction

An overarching theme of this thesis is to question the assumption that there is a single proxy to “ground truth” in the assessment of mindreading. One way to examine this question is by exploring if task-specific factors change the “best” description perceived by participants. Study 4 found that variation in task format influenced preferences for mental state interpretations. In the current chapter, the overarching aim is to examine whether mental state interpretations of a target is influenced by information about the context in which the target is engaging in a social interaction. Individuals might also vary in their flexibility to adjust their mental state interpretations with reference to the context of social interaction. This is relevant to another overarching aim in the current thesis: to explore potential indices of individual differences in adults’ mindreading abilities.

In Study 3, an open-ended task was presented to examine if participants showed consistent individual differences in their propensity to generate multiple interpretations and in aligning with other participants’ interpretations across items. However, no reliable individual differences in the latter were observed. One possible reason is that alignment on selection of the most plausible interpretation is not a stable trait, but the absence of within-person stability in alignment can also be due to methodological limitations in the task design. Specifically, the freedom to generate and select interpretations in the open-ended format and the minimal contextual constraints might make it more challenging to align with others. If this is the case, restricting the range of possible interpretations might result in higher inter-item correlation of alignment scores. A forced-choice version of the task was developed in Study 4, and the current studies were built on the forced-choice task. A large sample recruited for Study 7 made it possible to investigate inter-item correlations of alignment scores when context was absent or introduced.

5.1.1 Effect of context

Context could influence social perception and reasoning through the activation of schemas. Schemas are cognitive structures that organise knowledge about specific objects or events in a structured manner, and scripts are a specific kind of schema that describe the sequence of generalised actions in a given event or context (Cantor et al., 1982; Schank & Abelson, 1977; Taylor et al., 2023). These cognitive structures help individuals navigate through daily social interactions and enable coordinated behaviour in social situations.

Research has suggested that social differences among neurodivergent individuals were associated with difficulties in generalising and sequencing social events as scripts (Loth et al., 2008). Neuroscience studies have demonstrated increased brain synchrony when context was provided (Hasson et al., 2012), which provides a plausible neural mechanism for explaining the influence of context in individuals' social interactions with others. Relevant social cognition research on emotion perception has also shown that perception of emotions through facial expressions was influenced by contextual information, including but not limited to descriptions of social situations (Barrett et al., 2011; Carroll & Russell, 1996). Hence, it is reasonable to suspect that context plays a role in the process of mindreading by activating social scripts to aid one's interpretations of others' mental states and guide one's action in social situations.

Nevertheless, it should be noted that mindreading is likely more than just reading social scripts, as a generalised social script unlikely encompasses all variations in similar scenarios, for example, inferring whether the person with whom one is interacting is being sarcastic or truly convinced of what they are saying (Apperly et al., 2024). This idea is also consistent with the simulation account of mindreading, which suggests that individuals can understand others mental states by simulating others' mind with one's own mind; one's past experience in social situations is likely to influence the simulation process, but the reasoning is not solely dependent on social scripts (e.g., Goldman, 2006; Harris, 1992).

Despite the plausible influence of context on mental state attributions, some mindreading tasks, notably mental state decoding tasks, are decontextualised. A representative example is the RMET, which has been identified as the most frequently used measure in adults in Chapter 2. If context biases people's mental state interpretations, it might be inappropriate to assume the model answer is always by default the best description regardless of context. This idea is consistent with previous criticism of the conventional false belief task, which arguably does not require children to consider social relationships between characters or other social information (Killen et al., 2011). There has been a long line of work investigating whether manipulating factors such as motive of the character transforming the target object in false belief tasks influenced children's performance (Wellman et al., 2001). Furthermore, more current studies of mindreading have shown that even adult participants were less inclined to choose the default answer in the change-of-location false belief tasks when provided with more, particularly inconsistent, information about the target (Cho et al., 2022), suggesting the provision of contextual information influences mindreading decisions.

Context can be broad or specific. A specific type of context could be characteristics of the target person whose mental states are to be interpreted. Relevant studies of the mind-space theory have demonstrated that personality traits attributed to a target influenced interpretations of the targets' mental states, and participants updated their interpretations of the target's mental states when the target's behaviour mismatched the information about their personality traits provided by experimenters (Conway et al., 2019; Long et al., 2022). However, a limitation of the mind-space studies was that they focussed on how people located a target's mind in a structure of personality dimensions, while the broader context of social interactions has not been scrutinised. Apart from work on the mind-space theory, other research has shown that an individual's familiarity with the target influences the individual's mindreading performance (Zaki et al., 2009). Similarly, in studies of children mindreading, research has shown that children performed better when reasoning about the mental states of targets from the same cultural background than targets from other cultural background (Gönültaş et al., 2020; Perez-Zapata et al., 2016). Furthermore, both children and adults were found to take the action history of a target person into consideration when predicting how the target would think, feel and act, and the coherence among these three components (Lagattuta et al., 2016). These findings all suggest that the mindreader's perception and knowledge about the target, which can be seen as a specific type of context, influence mental state attributions to the target.

The general impact of the broader context of social scenarios has been sparsely investigated in the field of adult mindreading research. In philosophical work of mindreading, Spaulding (2018) suggested that the existing literature tends to overlook the significance of context for mindreading and called for more empirical work on assessing the influence of context on mindreading. There are also more extreme theories that suggest script reading, rather than mindreading, is central to explaining and predicting behaviour (e.g., Eickers, 2024; Taylor, 2023). With reference to parallel research discussed in the paragraphs above, it is reasonable to anticipate that contextual information influences interpretations of mental states, particularly in ambiguous social situations. If the effect of context is observed, the observation would provide further justification for using naturalistic tasks such as the MASC, which are designed to include numerous cues that indicate the context of social interactions.

5.1.2 Flexibility to context as a potential source of individual difference

Another overarching theme of this thesis is to explore indicators of individual differences in neurotypical adults' mindreading, and flexibility has been suggested to be a potential source of individual differences of neurotypical adults' mindreading (Chapter 1).

Recent theories suggest that the ability to interpret social situations flexibly by integrating contextual information into mental state interpretations is crucial for mindreading (Apperly et al., 2024; Devine, 2021; Hughes & Devine, 2015), and individuals are likely to differ in their abilities to integrate context into their interpretations when faced with varying contexts. One way to test this is to expose participants to a diverse set of contexts and observe whether they perform well in interpreting the characters' mental states across all contexts. This approach captures an individuals' ability to integrate a static context in their mental state interpretations. Another way to test this is to evaluate participants' tendencies to adjust their mental state interpretations of the same target when the context changes, for example, by reducing the plausibility of their previous judgment and suggesting an alternative interpretation through specific manipulations. This approach captures a dynamic process of adjustment with reference to context. While most existing mindreading tasks focus on the former approach, the studies presented in this chapter introduce a novel approach by evaluating dynamic context adjustments.

Regardless of the approach adopted, assessing flexibility in empirical studies involves evaluating the appropriateness of changes in interpretation, as not all changes reflect appropriate context integration. This characteristic distinguishes flexibility from the propensity to mindread. For instance, an interpretation viewed as highly plausible by many people in a specific context might be perceived as less likely in another context. If an individual perceives the interpretation to be more likely in the second context than in the first context, it tends to be considered inappropriate with reference to the preference of the majority. In the present studies, it was expected that the provided context information would suggest a specific interpretation. However, this needed to be verified empirically by checking whether participants showed an increased tendency to endorse the intended interpretation. Such an increased tendency among participants would justify that assigning a higher plausibility ranking to the target interpretation was appropriate, thereby providing a baseline for differentiating between desirable and undesirable shifts in participants' rankings of interpretations according to their plausibility in varying contexts. Specifically, individuals were expected to be capable of adjusting their interpretations of a target's thoughts or feelings when the context suggested a specific interpretation that differed from the most popular interpretation when no context was provided. Variability in people's likelihood of adjusting their interpretations in response to context changes, which characterises flexibility, was also expected.

Moreover, Study 4 demonstrated a typically chosen interpretation for each item when no additional context was given. This allowed for directly adopting the options from Study 4 to create contexts suggesting specific interpretations, bypassing the need to consider a huge variety of less likely interpretations when generating the contexts. It was hypothesised that participants would be more inclined to endorse the target interpretation suggested by specific contexts than when no context is given.

Once it had been established that participants could adjust their interpretations of a target's mental states with reference to context in the way intended by the manipulation, participants' flexibility can be evaluated by counting the number of times they assigned a higher plausibility ranking to an interpretation when it was suggested by the context compared to the baseline condition when no context was provided. Variations in participants' performance in making such adjustments can be scrutinised to determine the presence of reliable individual differences in participants' tendencies to adjust interpretations by taking context into consideration, which characterises flexibility. Consistent individual differences would be evidenced by satisfactory inter-item correlations.

5.1.3 Outline of studies in the current chapter

The studies in this chapter focus on the effect of context. The tasks involve presenting information about the relationships among characters or the settings of their interactions as contextual information alongside the pictorial stimuli.

Study 5 was comprised of a series of four between-participants experiments testing whether participants' interpretations of targets' mental states were influenced by the manipulated contexts. The dependent variable was the ranking of interpretations based on perceived plausibility, a modification from the single forced-choice format of Study 4.

Study 6 employed a within-participant design to investigate whether participants would alter their judgment of the most plausible interpretation upon receiving contextual information, and if they could shift their interpretations back and forth while still differing from the baseline condition as a demonstration of flexibility. However, due to the small sample size in Study 6, inter-item correlations were not calculated.

Study 7 sought to replicate and extend the findings of Study 6 with a sample large enough to examine individual differences.

In the subsequent sections, the methods and results of each study will be discussed individually, followed by an overarching discussion synthesising the conclusions drawn from Studies 5 to 7.

5.2. Studies 5a, 5b, 5c, and 5d: Between-participant design and stimulus refinement

Study 5 consisted of four small-scale experiments. For brevity these studies will be reported in a single method and results section. The overarching aim was to investigate whether contextual information could systematically alter interpretations of mental states in ambiguous social stimuli. An option that few participants endorsed as the most plausible interpretation (i.e., the low-frequency option) and another option that many participants selected as the most plausible interpretation (i.e., the high-frequency option), were identified from Study 4 as target interpretations for each picture.

It was predicted that participants would be more likely to rank the target interpretation as the most plausible when the corresponding context was provided, compared with the baseline condition where no context was presented. This effect was predicted to be stronger for contexts suggesting low-frequency interpretations. This is because for high-frequency interpretations, the target interpretations were already likely to be endorsed by participants even without contextual information, leaving not much room for further enhancement, while there was much more room for increasing endorsement of low-frequency interpretations. Additionally, it was expected that providing information about context would lead to higher overall alignment among participants in their interpretations. As alignment was calculated as the proportion of participants agreeing with each others' choices, through narrowing down the set of candidate interpretations participants would consider by providing context, alignment was likely to be enhanced.

Furthermore, this series of studies served as pilot studies for the subsequent study (Study 7) on flexibility. Depending on the findings, the context information and interpretation text options were refined to influence interpretations more effectively, as this would help develop better stimuli for assessing shifts in perceived plausibility that align with other participants in subsequent studies that target on measuring flexibility.

5.2.1 Method

Study 5a was preregistered on OSF before data collection (https://osf.io/w7jhv/). Studies 5b-d were follow-up studies to refine the stimuli, contextual information, and alternative interpretations based on preliminary findings from their previous studies.

Participants. Sixty participants with balanced gender were recruited for each of the four studies, which was the sample size needed for having a minimum frequency of five in each cell of a χ^2 test of independence with 12 conditions (3 groups * 4 options for each item). All participants were aged between 18 and 25 years (Range of M_{age} = 22.50 to 23.00 across four studies). In each study, half (n = 30) of the participants were female. All participants were recruited online via Prolific with the following screening criteria: UK residence, speaking English as their first language, had not been diagnosed with ASD, and had not participated in the previous series of studies. Most participants were monolingual (78.3% to 83.3%), while fewer participants were bilingual (13.3% to 20%), and few were multilingual (1.7% to 5.0%). Slightly over half of the participants held a Bachelor's degree or above (55% to 65%). Most participants identified as White (n ranged from 39 to 46, or 65.0% to 76.7%), followed by Asian (n ranged from 6 to 13, or 10.0% to 21.7%), Mixed (n ranged from 3 to 6, or 5.0% to 10.0%), Black (n ranged from 2 to 4, or 3.3% to 6.7%), and Arab (n =1 or 1.7% from Study 5d).

Study design and procedure. All four studies followed a between-participants design. Participants were randomly allocated into three groups (n=20 in each group), including two experimental conditions (i.e., high-frequency context group or low-frequency context group) where participants were shown contextual information, and a baseline control condition in which participants were not given any contextual information. In the high-frequency context condition, participants read contexts suggesting an interpretation that were among the top two most popular for each item from Study 4. Conversely, in the low-frequency context condition, participants read context suggesting less frequent interpretations, specifically within the two least popular interpretations from Study 4.

In each study, all participants gave informed written consent approved by the Ethical Review Committee at the University of Birmingham before participation. They completed an online questionnaire on the Qualtrics survey platform, in which they were shown five (Study 5d) or six (Study 5a-c) pictures depicting naturalistic social scenarios, the same stimuli used in Study 2 to 4. The presentation of the pictures as well as the order of the alternatives were randomised.

In the experimental groups, participants were instructed to read a sentence describing either the relationship between characters or the general setting (e.g., “They are a couple” or “They’ve been told that an incident has occurred inside”) before evaluating the plausibility of four alternative interpretations. In Study 5a, the sentence describing context was presented above each picture, bolded, in font size 22. In Studies 5b-d, compulsory timers were added such that participants were forced to stay on the same page for six seconds to read each context sentence before the picture appeared below the sentence. In contrast, the baseline group was only instructed to evaluate the plausibility of four alternative interpretations by referring to the pictures. No text was presented alongside the pictures.

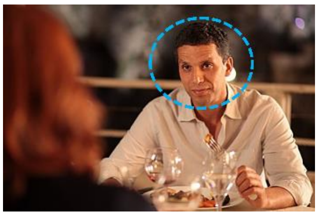
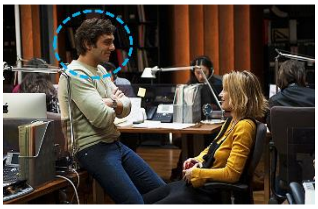
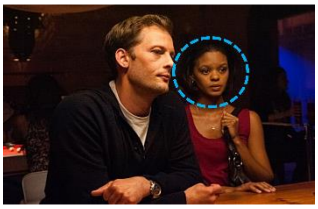
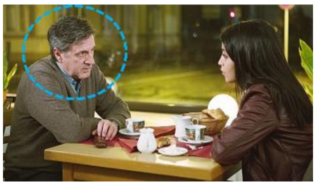
Unlike Study 4, where participants had to select one out of four plausible interpretations, participants in Studies 5a-d were required to rank the interpretations from the most to the least plausible by dragging them to rearrange the order. The instructions stressed that there was no definite right answer. Testing was completed in one session and the duration of the session was around five minutes. Participants received £0.75 for completing the study.

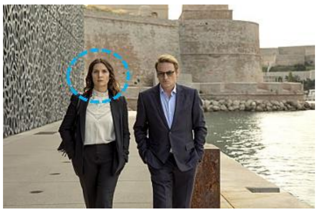
Materials. Studies 5a-c used the same six pictures (P3, P4, P5, P8, P9, P10) as in Studies 2 to 4. In Study 5d, P4 was dropped as the context suggesting the low-frequency target interpretation was ineffective despite attempts at refinement.

Alternative options. In Study 5a-b, the same four alternative interpretations used in Study 4 were presented. In Study 5c, the interpretation text for three out of four options (including the two target interpretations in the experimental conditions) were re-selected from candidate entries (i.e., verbatim responses from participants from Study 3) to further reduce similarity between the different interpretations. It was ensured that the re-selected interpretations were coded the same categories as the original options. In Study 5d, the text for the two non-target interpretations were slightly modified to reduce similarity with the low-frequency target interpretation in P3. The category chosen as the low-frequency target interpretation for P5 was changed from “contemplating” to “proud/arrogant” because the original option for “contemplating” was almost not endorsed by any participants across Studies 5a-c, either with or without contextual information presented. A verbatim response coded as “proud/arrogant” from Study 3 was used as the new alternative interpretation.

Context. The context information presented in the two experimental groups was created by the research team (Yeung, Devine, Apperly), each intending to suggest a specific interpretation of the target characters’ mental states. They were crafted to favour one interpretation over others without completely ruling out alternative options, preventing participants from ranking interpretations solely based on logical deduction. In Study 5a, the first version of context information comprised short, simple sentences briefly describing either the characters’ relationship or the interaction’s setting (e.g., “They are colleagues” or “They are on their way to a meeting”). In Study 5b, context sentences were elaborated to provide more details on the relationship and setting for each picture, maintaining one sentence per stimulus for each experimental group (e.g., “The two colleagues are having dinner together after a work meeting”). In Study 5c, the contexts for P5 and P8 were further modified. In Study 5d, the context for the low-frequency condition for P5 was rewritten to align with the changed interpretation category. Table 5.1 shows the list of alternative interpretations and the context information presented in Study 5d. The list of options and context presented in Studies 5a-c are available in Appendix C.

Table 5.1. List of context and interpretations (options) presented in Study 5d.

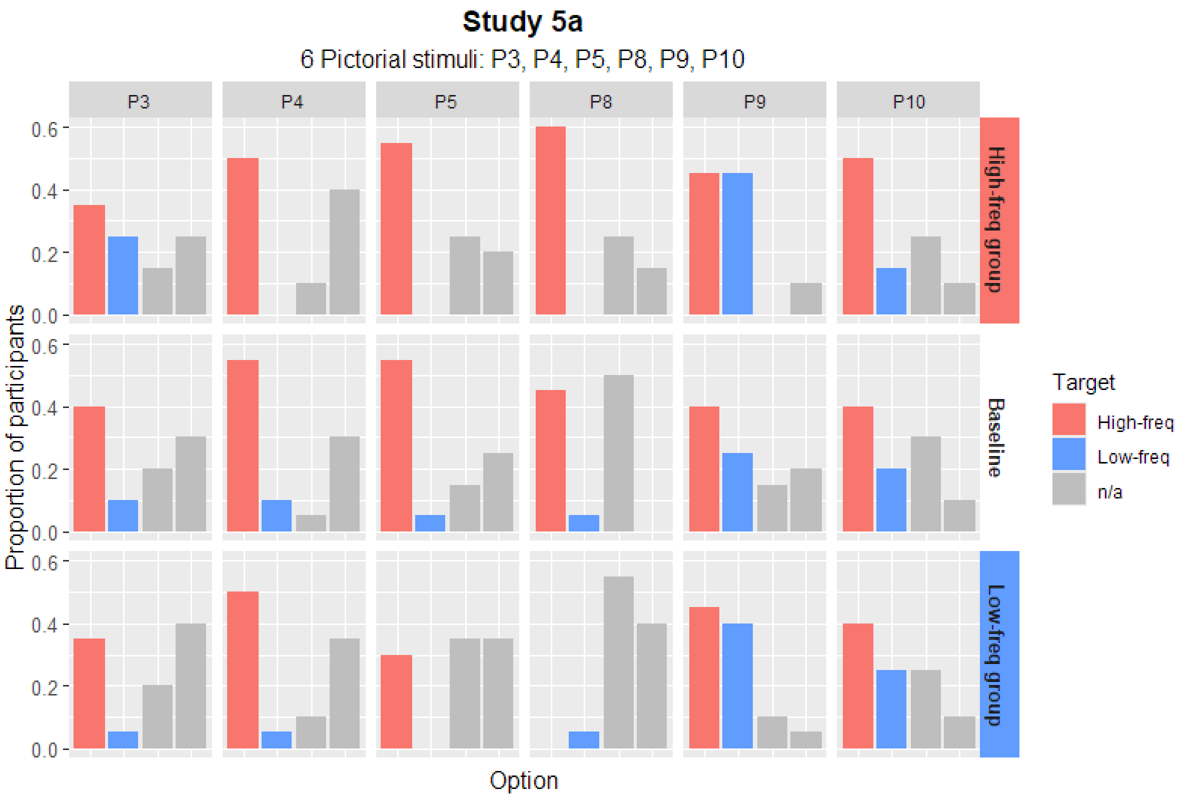
Picture	High-frequency context	Low-frequency context	Interpretations
P3 	The two colleagues are having dinner together after a work meeting.	The couple is sharing a meal on their anniversary.	He is interested in what she is saying. (High-frequency target)
			He is in love with his partner. (Low-frequency target)
			He is wondering whether this meal is worth it.
			He is feeling happy.
P5 	The colleagues have found that they are both free after work today.	He just got promoted to leader of their team.	He is feeling amused.
			He feels attracted to her. (High-frequency target)
			He is smug and self-satisfied. (Low-frequency target)
			He is feeling relaxed.
P8 	They've been told that it hasn't yet been possible to contact their daughter.	They've been told that their reservation was cancelled.	She is feeling anxious. (High-frequency target)
			She is feeling shocked.
			She is very sad about something somebody has said.
			She is feeling hugely annoyed. (Low-frequency target)
P9 	They are meeting because his daughter called him.	He just invited his daughter out to tell her his decision to divorce her mother.	He is focused on what she is saying. (High-frequency target)
			He is worried about the news he is about to pass on. (Low-frequency target)
			He is angry and annoyed about her attitude.

			He is feeling disappointed with his daughter.
P10 	They are on the way to handling a difficult assignment	The couple is on their way home from lunch.	She is feeling angry with him. (Low-frequency target)
			She is concentrating on an upcoming meeting.
			She is problem solving.
			She is determined and about to take on a challenge. (High-frequency target)

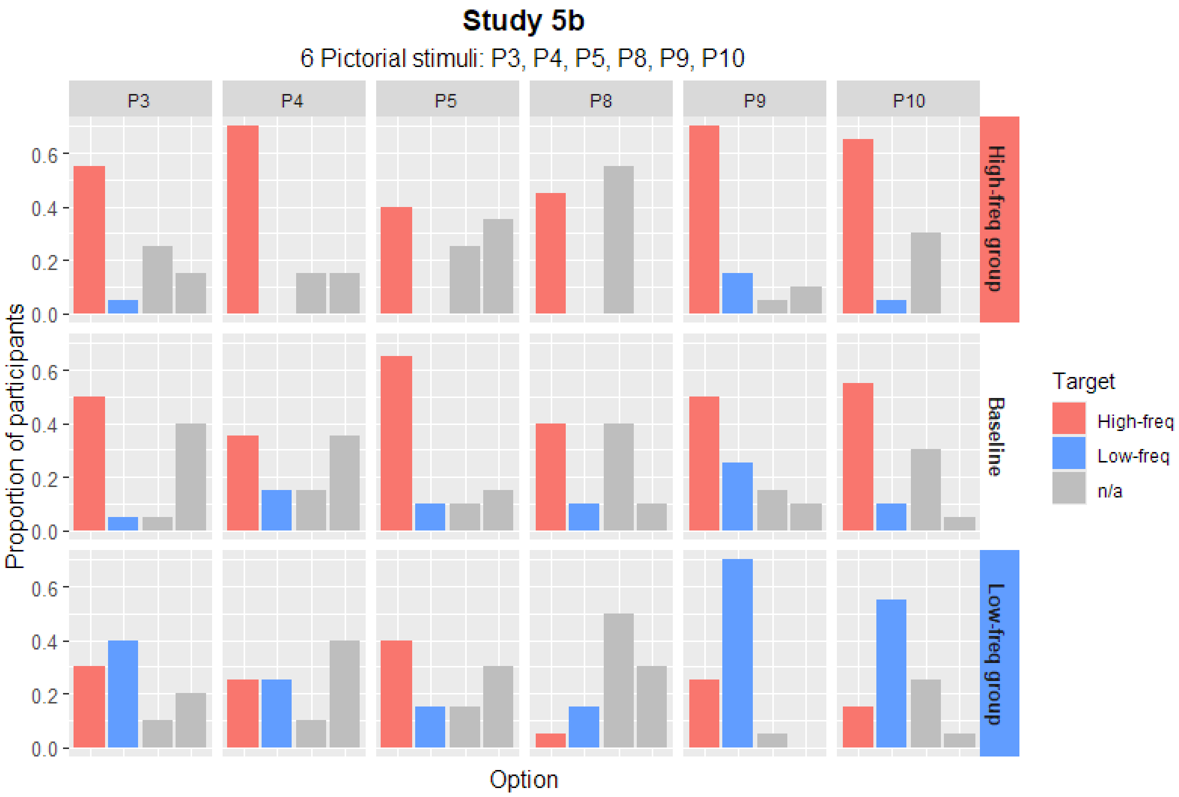
5.2.2 Results

The numbers of participants ranking the four alternative interpretations as the most plausible for each picture in the three groups are presented in Figure 5.1. If context effects occured, then the high-frequency contexts (top panels) would lead to more participants endorsing high-frequency target options, and the low-frequency contexts (bottom panels) would lead to more participants endorsing low-frequency target options. The baseline group provide a reference for comparison with each experimental group.

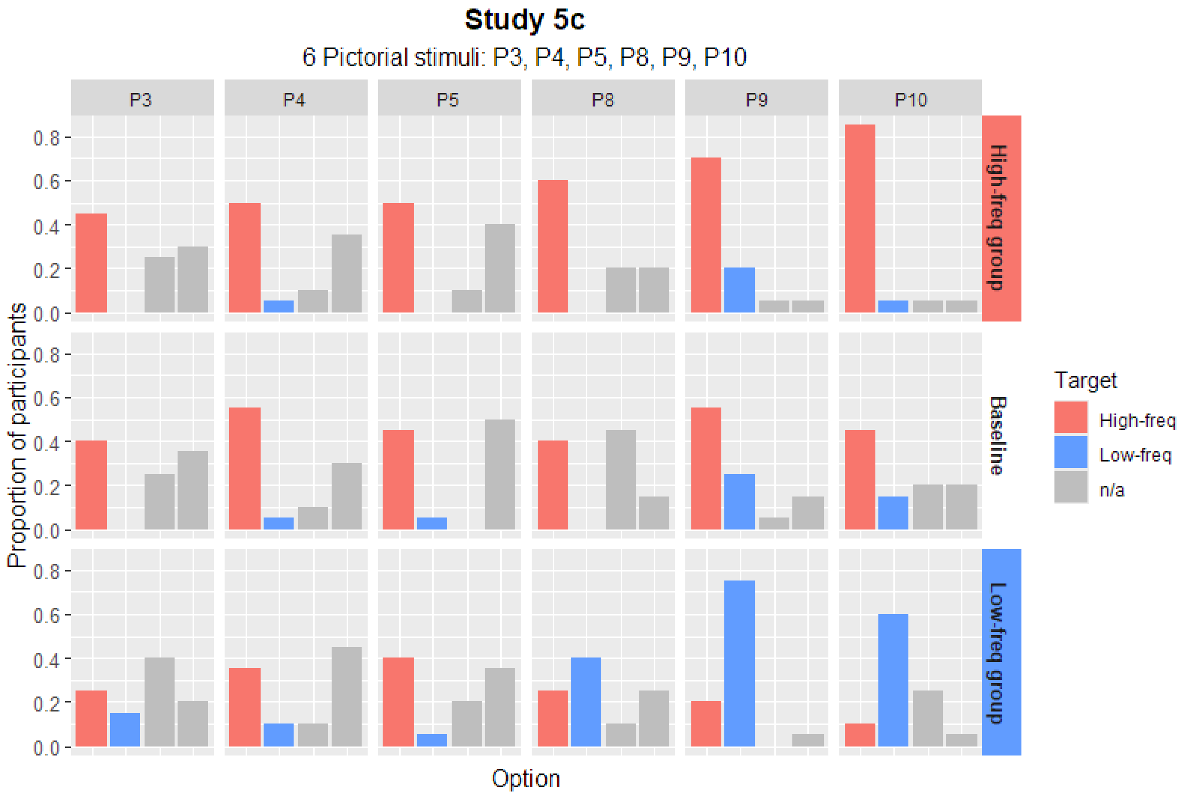
(a)


Study 5a
6 Pictorial stimuli: P3, P4, P5, P8, P9, P10
P3
P4
P5
P8
P9
P10
High-freq group
Baseline
Low-freq group
Proportion of participants
0.0
0.2
0.4
0.6
Option
Target
High-freq
Low-freq
n/a

(b)


Study 5b
6 Pictorial stimuli: P3, P4, P5, P8, P9, P10
P3
P4
P5
P8
P9
P10
High-freq group
Baseline
Low-freq group
Proportion of participants
0.0
0.2
0.4
0.6
Option
Target
High-freq
Low-freq
n/a

(c)


Study 5c
6 Pictorial stimuli: P3, P4, P5, P8, P9, P10
P3
P4
P5
P8
P9
P10
High-freq group
Baseline
Low-freq group
Proportion of participants
0.0
0.2
0.4
0.6
0.8
Option
Target
High-freq
Low-freq
n/a

(d)

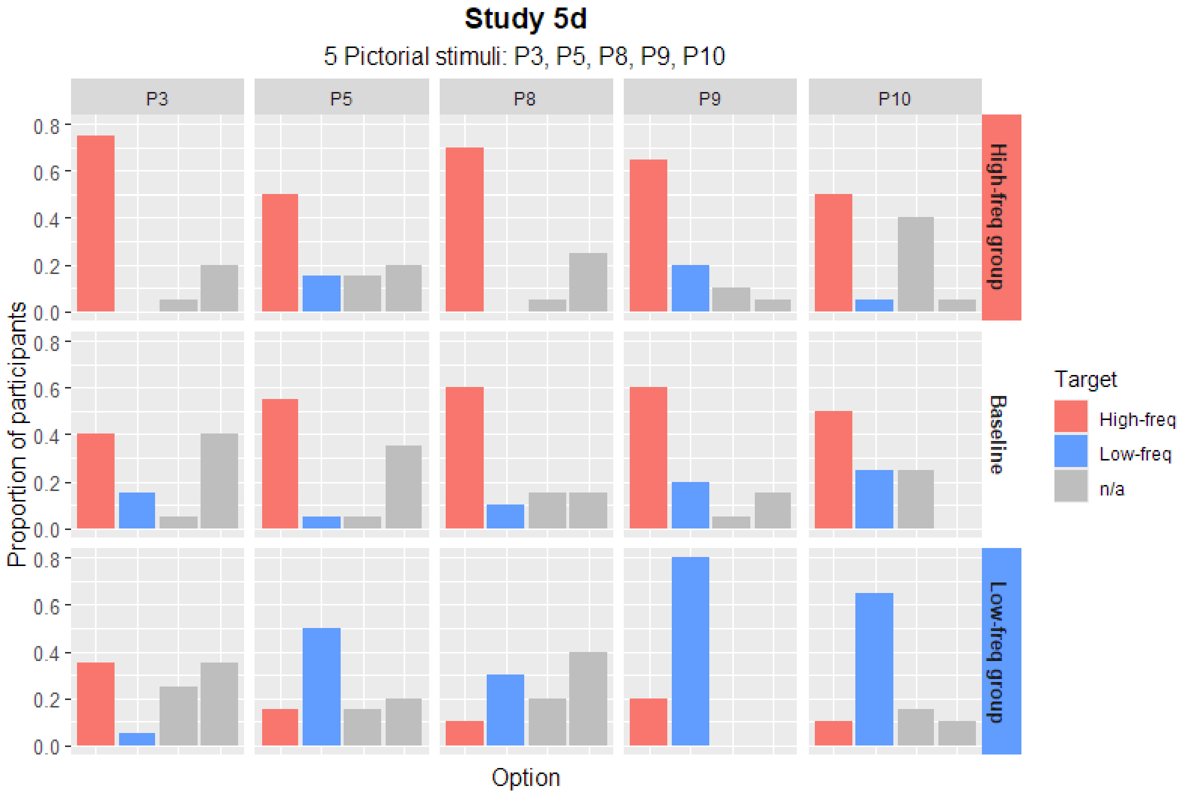

Study 5d
5 Pictorial stimuli: P3, P5, P8, P9, P10
P3
P5
P8
P9
P10
High-freq group
Baseline
Low-freq group
Proportion of participants
0.0
0.2
0.4
0.6
0.8
Option
Target
High-freq
Low-freq
n/a

Figure 5.1. With stimulus refinement from Study 5b-d, generally more participants ranked the target alternative interpretations as the most plausible, especially for the low-frequency targets. Stimulus P4 was dropped in Study 5d.

Overall effect of context on interpretation. Two separate Mann-Whitney U tests (due to violation of normality assumption) were conducted to compare how frequently participants in each experimental group and the baseline group ranked the target interpretation as most plausible, for each study. It was predicted that when context was given, participants would be more likely to rank the target interpretation first, in comparison to when there was no context, especially for the comparison between the low-frequency group and the baseline group. The results are summarised in Table 5.2.

Table 5.2. Summary of results in comparing the frequency of selecting target interpretations between experimental groups and the baseline group in each study.

	Frequency of hitting low-frequency (LF) targets			**Frequency of hitting high-frequency (HF) targets**		
Study	LF group *Mdn*	Baseline group *Mdn*	Mann-Whiteney test result	HF group *Mdn*	Baseline group *Mdn*	Mann-Whiteney test result
5a	1	1	U = 210.5, p = .770, r_g = .053	3	3	U = 219, p = .603, r_g= .095
5b	2	0.5	U = 337.5, p < .001***, r_g = .688	3	3	U=258.5, p = .097, r_g = .293
5c	2	0	U = 331, p < .001***, r_g = .655	4	3	U=281.5, p = .022*, r_g = .407
5d	2	1	U = 356, p < .001***, r_g = .780	3	3	U = 237, p = .300, r_g = .185

* p < .05

Note. The Glass rank biserial coefficient (r_g) is reported as an effect size measure for the Mann-Whitney tests (Mangiafico, 2022).

In Study 5a, Mann-Whitney U tests showed no significant differences between the experimental and baseline groups in ranking the target interpretations as most plausible. However, the lack of effect could have been due to the context information not being sufficiently salient in presentation such that participants did not pay attention to them, and/or being too simplistic to influence interpretations.

Following procedural and contextual modifications described in the methods section, a significant context effect was observed throughout Study 5b-d, as participants more often ranked the target low-frequency interpretations as the most plausible in the low-frequency experimental group than in the baseline group. In Study 5c, participants in the high-frequency experimental group were also more likely to rank high-frequency interpretations first than participants in the baseline group.

Narrowing down interpretations. To examine if participants narrowed down the candidate interpretations with context presented, alignment scores (i.e., the proportion of other participants who ranked the same interpretation as the most plausible for the same item) were also calculated for each item and averaged across items for each participant. The average alignment scores were then compared across the three groups using Kruskal-Wallis rank sum tests as the normality assumption was violated.

If more participants agreed with one another for the items, average alignment score would be higher. Higher average alignment scores were predicted in the two experimental groups than in the baseline group. Where an overall effect in average alignment difference was significant, additional Mann-Whitney U tests were conducted to compare the average alignment scores between each experimental group and the baseline group.

Kruskal-Wallis tests indicated no significant differences in alignment scores across the groups in Study 5a, χ^2 (2) = 1.79, *p* = .409. After refining the stimuli, test results for Studies 5b-d all showed significant differences, χ^2 (2) = 20.29 (Study 5b)/21.25 (Studies 5c-d), all *ps*

< .001. Follow-up Mann-Whitney U tests in Studies 5b-d consistently showed higher alignment scores in the high-frequency group compared to the baseline group. However, there was no significant difference in alignment scores between the low-frequency and baseline groups. The medians of the alignment scores in each group and Mann-Whitney U test results are reported in Table 5.3.

Table 5.3. Median of participants' average alignment scores in each group and Mann-Whiteney U test results.

Study	Baseline group median	Low-frequency (LF)/ High-frequency (HF) context group median		Mann-Whitney U test
5a	0.325	LF	0.342	n/a
		HF	0.346	
5b	0.338	LF	0.342	U = 210.5, p = .786, r_g = .053
		HF	0.461	U = 60, p < .001***, r_g = .700
5c	0.36	LF	0.351	U = 187.5, p = .745, r_g = .063
		HF	0.443	U = 44, p < .001***, r_g = .780
5d	0.374	LF	0.389	U = 175.5, p = .516, r_g = .123
		HF	0.463	U = 105.5, p = .011*, r_g = .473

* p < .05

*** p < .001

5.2.3 Discussion

Studies 5a-d aimed to examine whether providing context effectively influenced participants' mental state interpretations, especially when the context presented suggested a low-frequency interpretation. In summary, following modifications in stimuli and procedures, Studies 5b-d presented robust evidence that when the context suggested a low-frequency interpretation, participants were more inclined to rank the target interpretation as the most plausible compared to the baseline group. However, the comparison between the high-frequency context group and the baseline group had inconsistent results. This was likely due to the fact that the high-frequency interpretations were already popular in the baseline condition, as found in Study 4. The observed difference between the low-frequency group and the baseline group supported significant influence of context on interpretation; provision of context information effectively influenced participants' judgments about the most plausible interpretation of a character's thoughts or feelings in a given scenario.

Although there was no significant difference in the frequency of ranking the high-frequency interpretation as the most plausible in the high-frequency context group compared to the baseline group, Studies 5b-d consistently showed higher average alignment scores in the high-frequency context group than in the baseline group. This finding suggests an effect of context in another way: the distribution of responses became less dispersed among alternative options when a relevant context was introduced, even though the endorsement of the high-frequency target response did not increase overall. Hence, the effect of context was evident in both the low-frequency and high-frequency groups when compared to the baseline group, although manifested differently in each experimental group.

To follow up, a study design that more sensitively captures the context effect was required. The design should allow for examining changes in the ranking of interpretations, as an increase in perceived plausibility does not necessitate placing an option as the most plausible interpretation. Instead, an upward shift in plausibility ranking sufficiently would mark one's incorporation of context into mindreading interpretations. Therefore, a within-participant design was adopted in Study 6. The within-participant design was also crucial for examining flexibility dynamically, as shifts in perceived plausibility of various interpretations based on given context is inherently within-person.

5.3. Study 6: Within-participant design

Study 6 aimed to explore the effect of context on interpretation of a target's mental states with a within-participant design. This study also specifically examined whether individuals shifted their perceived most plausible interpretation in response to contexts that suggested different interpretations. Any variation among participants' tendencies to alter plausibility rankings, specifically in the intended direction suggested by the context, was also examined to establish a basis for calculating flexibility in the subsequent study (Study 7). Based on Study 5, it was predicted that (1) participants would be more likely to rank the target interpretations as the most plausible when the corresponding contexts were presented in comparison to the baseline condition; (2) the distribution of participants' total frequency of altering rankings of the target interpretations in the intended direction should be different from a uniform distribution.

There could be a potential sequence effect as well: exposure to low-frequency context prior to high-frequency context, or vice versa, might influence participants' response patterns. It is important to rule out this sequence effect for future studies focused on flexibility.

5.3.1 Method

Participants. A priori power analysis indicated that 34 participants were required to detect a medium effect size of Cohen's d = 0.5 with .80 power at α = .05 in a paired t-test[1] using G*Power (Faul et al., 2009), following a similar rationale to the effect size specified for Study 2 (Chapter 3). To balance gender in the two counterbalanced versions of the task, 36 participants aged between 18 and 25 (18 female, M_{age} = 22.42) were recruited for Study 6. All

[1] The initial power calculation failed to take into account the need for a Bonferroni correction; 41 participants are required to detect a within-participant main effect of a moderate effect size (f=0.25) in a repeated-measures ANOVA with Bonferroni-corrected significance level at .025 at 80% power based on G* Power (Faul et al., 2009). However, Study 7 provides a replication of the effect of interest in a much larger sample.

participants were recruited online via Prolific with the same screening criteria as in the previous series of studies. Most participants were monolingual (77.8%), followed by bilingual (16.7%), and multilingual participants (5.6%). Slightly less than half of the participants had a Bachelor's degree or above (47.2%). Most participants identified their ethnicities as White (58.3%), followed by Mixed (19.4%), Asian (13.9%), and Black (8.3%).

Study design and procedures. Study 6 adopted a within-participant design. The main independent variable, contextual information, had three levels, as in Study 5: the baseline condition, the high-frequency context condition, and the low-frequency context condition. The set of context information text presented was adopted from Study 5d. A blocked design was adopted, in a way that participants were first presented with all five pictures (used in Study 5d) in the baseline condition, followed by either the high-frequency or low-frequency condition (counterbalanced between participants), then the other condition where context information was provided.

In each study, all participants gave informed written consent approved by the Ethical Review Committee at the University of Birmingham before participation. All participants completed an online questionnaire on the Qualtrics platform, in which they were shown three blocks of the same five pictures as in Study 5d, each depicting a naturalistic social scenario with two individuals. As in Study 5, the participants were required to rank the four options given, each describing one possible interpretation of what the target character was thinking or feeling, in descending plausibility each time a picture was presented.

As the participants answered the same ranking question three times for each picture, the instructions were modified before the second block and the third block. Before the second block began, an instruction block was presented. The instructions stated that participants would see the same pictures as what they just saw, but before each picture was presented, they would first see a sentence describing relevant background information. It was stressed that there was no definite right answer, and it did not matter if participants remembered their previous answers or wanted to give different answers from before; participants were instructed to give the answers that made the most sense with the background information provided. The same instruction block was presented again after participants finished the second block and before they started the third block.

Testing was completed in one session and the duration of the session was around ten minutes. Participants received £2.25 after completing the session.

Materials. The five pictures, the context information, and the alternative interpretations were all the same as the materials used in Study 5d.

5.3.2 Results and discussion

Condition comparison. Analyses were conducted to compare the number of times the target interpretations were ranked first between experimental conditions and the baseline condition. Two separate 2 (condition: experimental vs. baseline) x 2 (counterbalance order) mixed ANOVAs were conducted. The main effect of condition was the key result, as it would show whether participants tended to select the target option more often in the experimental group compared to the baseline group. Counterbalance order was included in the models to check any order effect. The Bonferroni correction was applied to correct for the repeated use of baseline data. A significant condition*order interaction would indicate that the data from participants allocated to the two order versions should not be combined for evaluating the effect of context. The descriptive statistics are summarised in Table 5.4.

Table 5.4. Descriptive statistics for the mean number of times participants ranked the intended interpretation first in each group by counterbalance order; (a) shows the comparison between mean number of times ranking the low-frequency target option first in the baseline condition and low-frequency condition, while (b) shows the comparison between mean number of times ranking the high-frequency option first in the baseline condition and high-frequency condition.

(a)

Order	**Baseline condition mean (*SD*)**	**Low-frequency condition mean (*SD*)**
Order 1	0.89 (0.90)	3.06 (0.87)
Order 2	1 (0.84)	3.06 (1.00)

(b)

Order	**Baseline condition mean (*SD*)**	**High-frequency condition mean (*SD*)**
Order 1	2.94 (1.26)	3.72 (0.96)
Order 2	2.28 (1.32)	3.33 (0.77)

In the first ANOVA, the two levels for condition were “low-frequency context condition” and “baseline condition”. The main effect of condition was significant, *F*(1, 34) = 114.98, *p* < .001, partial eta-squared = .772. The main effect of order was not significant, *F*(1, 34) = .059, *p* = .809, partial eta-squared = .002. The interaction was not significant either, *F*(1, 34) = .08, *p* = .780, partial eta-squared = .002. The second ANOVA, in which the two levels for condition were “high-frequency context condition” and “baseline condition”, showed similar results. The main effect of condition was significant, *F*(1, 34) = 11.42, *p*

= .002, partial eta-squared = .251, while the main effect of counterbalance order was not significant at Bonferroni-corrected significance level, *F*(1, 34) = 4.59, *p* = .039, partial eta-squared = .119. The interaction was not significant either, *F*(1, 34) = 0.26, *p* = .612, partial eta-squared = .008.

The results showed that the main effect of condition was significant in both ANOVA models, indicating that across the two counterbalanced versions, participants were more likely to rank the target high-frequency option first in the context condition, *M* = 3.53 times, than in the baseline condition, *M* = 2.61 times. They were also more likely to rank the low-frequency option first in the context condition, *M* = 3.06 times, than in the baseline condition, *M* = 0.94 times. These findings support the first prediction of the study, replicating the context effect observed in Study 5 but with a within-participant design.

Moreover, the order of presenting the two experimental conditions did not interfere with participants' tendency to rank the target options first compared with the baseline condition, justifying further analysis that collapsed data from the two groups of participants.

Preliminary investigation on flexibility as a potential index of individual differences. The above analyses focused on the number of times participants ranked the target options first, but the effect of context can also be demonstrated by a participant shifting the ranking of the target option up, without ranking it first. However, if the target option was already the first-ranked option in the baseline condition, there would be no room for it to improve. Therefore, further analysis was conducted at an item level, collapsing the two experimental conditions: 1 mark was scored on each item if the participant shifted the ranking of a target option up at least once across the two experimental conditions compared to the baseline condition. Hence, the score, which measured flexibility, could vary from 0 to 5. The observed distribution was compared to a uniform distribution to examine whether there was variation in participants' flexibility by conducting a Kolmogorov-Smirnov test and plotting a histogram for visual inspection.

The one-sample Kolmogorov-Smirnov test result showed that the score distribution differed significantly from a uniform distribution, $D = .57$, $p < .001$, implying the existence of variation among participants' flexibility scores. Figure 5.2 shows the histogram depicting the distribution of participants' flexibility scores measured by the number of items in which they shifted the target options in either or both of the experimental conditions. The skewness and kurtosis of the distribution was -.41 and 2.35, respectively, indicating that the distribution did not deviate much from a normal distribution. However, by visual inspection of the histogram, the distribution tended to be skewed to the left. Moreover, participants' mean score was 4.06 out of 5, with 27 (74%) participants scoring 4 or above, suggesting a possible ceiling effect. Thus, a larger sample size was required to examine the variation in participants' flexibility score indicated by the number of items in which they shifted the ranking of the target interpretations in the intended direction at least once.

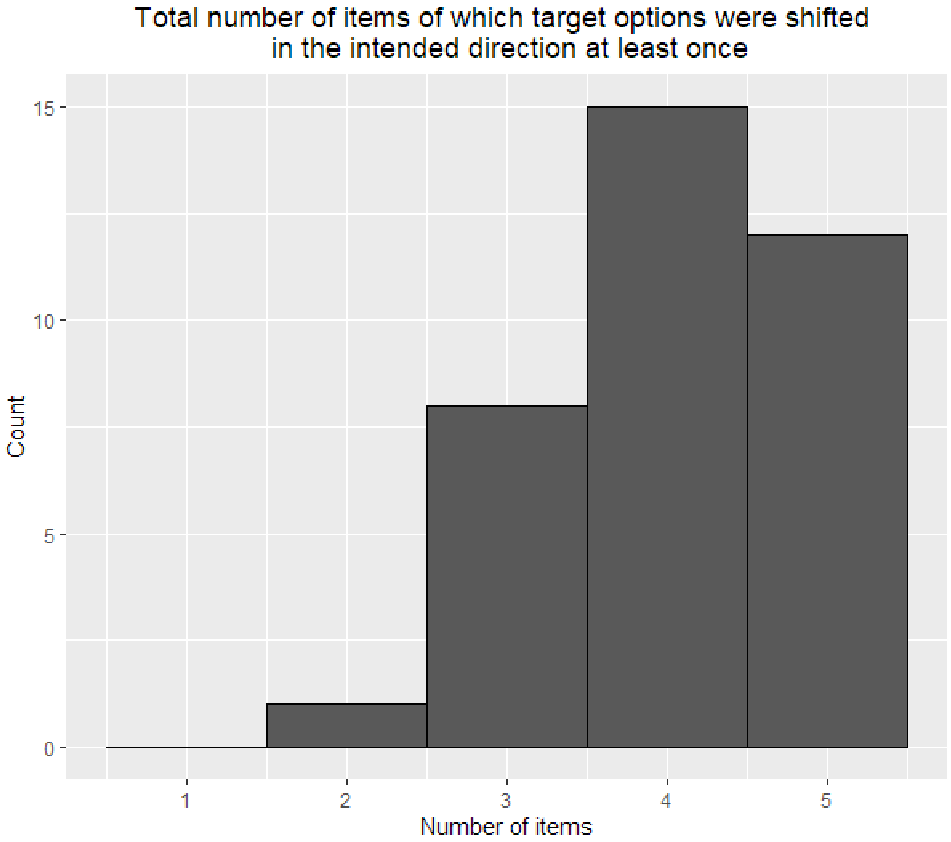


Figure 5.2. The histogram shows the distribution of the number of items of which participants shifted the target option(s) in the intended direction at least once, demonstrating a possible ceiling effect.

5.4. Study 7: Replication and examining individual differences

From Study 6, condition order did not alter participants' tendency to rank the target options first in comparison to the baseline condition. Hence, in the current study, the three conditions were presented in the same order across all participants to maximise detection of individual differences. Flexibility score was operationalised as the number of items (i.e., the pictures) in which participants shifted the target option in the intended direction in at least one experimental condition. To examine whether this score showed individual variation and whether the scores show good inter-item correlation, a larger sample size was recruited for Study 7. To summarise, the primary aims of Study 7 were to (1) replicate context effect in a within-participant design as in Study 6, (2) evaluate the presence of variability in participants' tendencies to shift the rankings of the target options in the intended directions with a larger sample size, and (3) evaluate the inter-item correlations of participants' flexibility scores on each item.

Furthermore, to address the question of whether the lack of inter-item correlations from Study 3 (Chapter 3) could be attributed to adopting the open-ended format and the lack of context, which could have allowed for greater variability of interpretations, the inter-item correlations of participants' scores on each item were inspected separately for the baseline condition and each of the two contexts.

5.4.1 Method

Participants. An a priori power analysis indicated that a sample size of 140 was required to detect factor loadings of .40, the conventional threshold for an acceptable factor loading, of five items loaded on a single latent variable with 80% power with a confirmatory factor analysis (CFA). Hence, 140 participants (70 females) aged between 18 and 25 (M_{age} = 22.56) were recruited online via Prolific. All participants were recruited online via Prolific with the same screening criteria as in the previous series of studies. Most participants were monolingual (67.9%), followed by bilingual (25.7%), and multilingual (6.4%). As for their educational background, 60% held a Bachelor's degree or above. Most participants identified their ethnicities as White (59.3%), followed by Asian (17.1%), Mixed (9.3%), Black (12.9%), Arab (7.1%) and others (7.1%).

Study design, procedures and materials. The study design, procedures and materials were identical to study 6 except that the presentation order of the two experimental blocks was not counterbalanced across participants to optimise the detection of individual differences. All participants first completed the baseline condition, followed by the low-frequency context condition, and finally the high-frequency context condition. Testing was completed in one session and the duration of the session was around ten minutes. Participants received £2.25 after completing the session.

5.4.2 Results and discussion

Replication of context effect. Results of paired t-tests showed that participants were more likely to rank the relevant target options first when high-frequency context information, *t*(139) = 5.16, *p* < .001, Cohen's *d* = 0.44, and low-frequency context information, *t*(139) = 18.23, *p* < .001, Cohen's *d* = 1.54, were provided than in the baseline condition, replicating the effect of context in Studies 5a-d and Study 6.

Variability in tendencies to shift the rankings of target options in the intended direction. As in Study 6, participants scored 1 point on each item if the participant shifted the ranking of a target option up at least once across the two experimental conditions compared to the baseline condition. Their scores on the five items were summed up as their total flexibility score. The observed distribution was compared to a uniform distribution with possible values from 0 to 5. The one-sample Kolmogorov-Smirnov test result showed that the score distribution differed significantly from a uniform distribution, *D* = .53, *p* < .001, implying the presence of variation in participants' flexibility scores. Figure 5.3 shows the histogram depicting the distribution of participants' flexibility scores measured by the number of items in which they shifted the target options in either or both of the experimental conditions. The skewness and kurtosis of the distribution was -.37 and 2.67, respectively, indicating that the distribution did not deviate much from a normal distribution; participants' mean score was 3.8 (*SD* = 0.90). However, by visual inspection of the histogram, the distribution tended to be skewed to the left. One-hundred-thirty (64%) participants scored 4 or above, suggesting a possible ceiling effect which could potentially affect the use of the score as an index of individual difference.

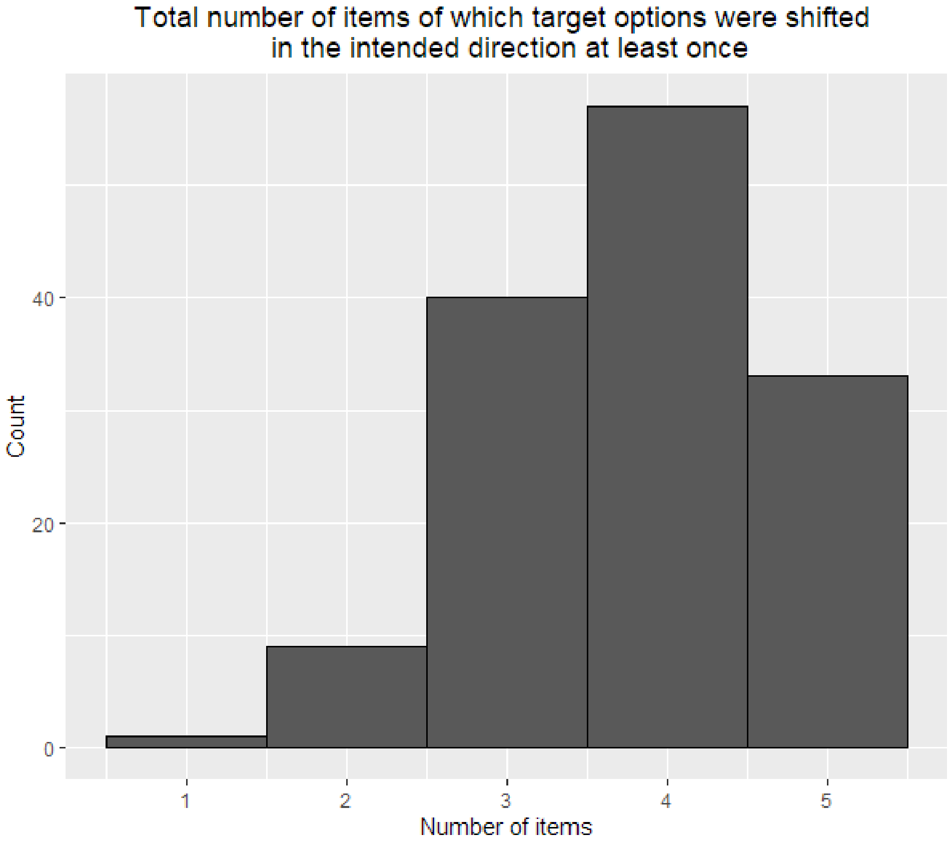


Figure 5.3. The histogram shows the distribution of the number of items of which participants shifted the target option(s) in the intended direction at least once, demonstrating a possible ceiling effect.

Evaluation of inter-item correlations in flexibility. Table 5.5 shows the tetrachoric correlations among participants' flexibility scores on each item. None of the positive correlations exceeded .30, while three correlations were negative. These results suggest that participants' flexibility scores did not correlate well among items, warranting no further need to inspect the items' loadings on a latent factor using CFA.

Table 5.5. Tetrachoric correlations among participants' flexibility score on each item.

	P3	**P5**	**P8**	**P9**
P5	.12			
P8	.04	.14		
P9	.07	.21	-.28	
P10	-.05	-.38	.17	.05

Evaluation of inter-item correlations in alignment. Extra analyses were conducted to separately evaluate whether inter-item correlations were satisfactory in terms of alignment in the three conditions, adopting the same .30 threshold as in Study 3 (Chapter 3). Alignment was calculated by the same method described in Study 4 and 5 with the current sample, by considering the proportion of other participants ranking the same interpretation as the most plausible as the participant concerned.

Baseline condition. Inter-item correlations in alignment were calculated in the baseline condition to target a question from study 3: whether a lack of inter-item correlation in alignment scores in an open-ended format of the task could be eliminated in a forced-choice format of the task, which limits the possible interpretations participants are allowed to choose from. Table 5.6 shows that none of the items were correlated in alignment scores.

Table 5.6. Spearman correlations among participants' alignment score on each item in the no-context condition.

	P3	**P5**	**P8**	**P9**
P5	-.15			
P8	.03	-.18		
P9	-.02	-.11	-.10	
P10	.03	-.04	.01	.02

Conditions with context. The same correlational analyses were conducted to examine participants' alignment with other participants when high-frequency contexts (see Table 5.7) and low-frequency contexts (see Table 5.8) were introduced. This was conducted to examine whether the inconsistent inter-item correlations in Study 3 could be eliminated when interpretations were further narrowed down by not only switching to a forced-choice format but also by introducing contextual constraints. Similar results with the baseline condition were found: negative correlations were often observed and positive correlations were weak, suggesting unsatisfactory inter-item correlations. None of the positive correlations exceeded the .30 threshold or were significant.

Table 5.7. Spearman correlations among participants' alignment score on each item in the high-frequency context condition.

	P3	**P5**	**P8**	**P9**
P5	-.05			
P8	-.02	.05		
P9	-.13	-.01	.03	
P10	-.07	-.06	-.12	-.29***

*** *p* < .001

Table 5.8. Spearman correlations among participants' alignment score, on each item in the low-frequency context condition.

	P3	**P5**	**P8**	**P9**
P5	.01			
P8	.063	-.05		
P9	-.08	.06	-.07	
P10	.06	.02	.15	-.17

To conclude, inter-item correlations were unsatisfactory in both flexibility and alignment. Participants' tendencies to adjust the rankings of interpretations was not an inappropriate indicator of individual differences in flexibility, which can be possibly due to the ceiling effect discussed. Moreover, switching to a forced-choice format or limiting the set of possible interpretations did not improve inter-item correlation in participants' alignment scores as an index of individual differences in mindreading performance.

5.5. General discussion

The over-arching aim of studies 5 to 7 was to investigate the effect of context on how people interpret others' thoughts and feelings in ambiguous social scenarios. Another aim was to explore the potential of using participants' tendencies to adjust their evaluations of possible interpretations, when provided with additional contextual information, as an indicator of individual differences in mindreading flexibility.

Context effect. Studies 5 to 7 showed that provision of information about the context of an ambiguous social scenario systematically altered participants' perceived best description of a target person's thoughts or feelings. These findings suggest that people do take context into consideration when trying to make sense of what others are thinking or feeling. This suggestion also aligns with recent studies in the mindreading literature showing that more available information about the target of mindreading influences one's inferences about the target's mental states (e.g., Cho et al., 2022; Conway et al., 2019). The current study attempted to evaluate the effect of context not limited to information about the target's personality or past behaviour. Furthermore, the current studies demonstrated that mental state inferences are influenced not only by introducing context about where a social scenario is taking place or the relationship between people in the interaction, but also that the "best" answer is likely an interpretation that varies depending on the context.

This finding might be partly explained by the function of social script, which is applied in comprehending unfamiliar social situations (Zacks, 2020): when information about the context of the interaction was presented, participants could make use of social scripts to make sense of the interaction and the mental states of the individuals involved in the interaction. However, the activation of social scripts did not necessarily rule out the alternative interpretations presented in the current task, as the alternatives were still plausible mental state interpretations. For example, given the context that "the couple is sharing a meal on their anniversary", the social script regarding sharing a meal on an anniversary does not rule out the possibility that the target "is interested in what [his dinner partner] is saying" or "wondering whether the meal is worth it". Hence, it was unlikely that participants only deduced the answers based on social scripts. Therefore, the current findings suggest that the incorporating context into mental state attribution is likely an element of the mindreading process on top of social script reading.

An implication of the current findings is that the "correct" answers for existing mindreading measures that present stimuli in a decontextualised manner might not hold true if contexts are introduced, and simple mental state "decoding" solely depending on observable cues might not be a sufficient explanation for mindreading success. If the attribution of mental states is influenced by context, the ability to integrate context into mental state interpretations might be at least as important as the ability to draw information from observable cues. This corroborates relevant research in emotion perception showing recognition of emotions from facial expressions is influenced by context, including but not limited to verbal descriptions of social situations; for example, when a story suggesting fear was presented with a facial expression of anger, participants were more likely to perceive that the stimulus was expressing fear (Carroll & Russell, 1996; for a review, see Barrett et al., 2011).

Lack of inter-item correlations. Despite the robust experimental effect, participants' tendency to make intended shifts was not a reliable indicator of individual differences in flexibility. This was possibly due to limited variance, as more than half of the participants scored on 4 or more items out of 5 items. Although the mean flexibility score was 76% (raw score: 3.8/5), falling below the 85% threshold set in the systematic review presented in Chapter 2, it should be noted that the current task only featured a limited set of five items. Restricted variance in participants' performance might have restricted inter-item correlations among items. The current findings on poor inter-item correlations also echo existing literature suggesting that experimental tasks are not always good tasks for measuring individual differences (Hedge et al., 2018). Another possibility for poor inter-rater reliability in flexibility is that adjusting one's evaluation of the plausibility of a specific interpretation of the ambiguous social scenario, in a way that is consistent with the majority, is not a trait-like tendency. However, this possibility warrants further research with designs in which ceiling effects are not a concern.

Additionally, alignment scores did not demonstrate satisfactory inter-item correlations. Hence, it was unlikely that the lack of inter-item correlation in Study 3 was due to the open-ended format allowing for greater freedom in participants' generation of possible interpretations than in forced-choice tasks or absence of contextual description.

5.5.1 Conclusion

In conclusion, the current findings suggest that context is important in interpreting others' thoughts and feelings in an ambiguous social scenario, but there is no conclusive evidence suggesting one's flexibility in adjusting interpretations along with changes in context can be a reliable index of individual differences. However, the lack of conclusion in the answer to the individual differences question can be due to methodological limitations of the present paradigm.

Chapter 6

Does context constrain both the generation and selection of interpretations?

6.1 Introduction

The focus of Chapter 5 was on the effect of context on participants' selection of the best mental state interpretation among given options. The current chapter extends the investigation by examining the role of context in both generating possible interpretations and selecting the best interpretations in mindreading. The aim of this chapter is to examine the extent to which context constrains the selection and generation of interpretations.

6.1.1 Generation and selection in mindreading

The idea of studying generation and selection as separate constructs in mindreading comes from both advances in recent research as well as long-standing theoretical models. As described in the previous chapters, recent research has drawn a distinction between the ability to infer mental states accurately and the propensity to make inferences or general social motivation (e.g., Devine & Apperly, 2022; Carpenter et al., 2016; Dodell-Feder et al., 2013). The distinction has not only been discussed theoretically but also investigated empirically. There is empirical evidence showing that ability and propensity are independent and predict different outcomes (e.g. Carpenter et al., 2016; Contreras-Huerta et al., 2020; Devine & Apperly, 2022; Lockwood et al., 2017). While accuracy typically involves comparing participants' mindreading responses to predetermined answers, a possible operationalisation of the propensity to make inferences is the tendency to generate multiple candidate interpretations of a target's mental states, regardless of their appropriateness. This operationalisation of propensity to mindread aligns with the suggestion that mindreading resembles adaptive reasoning, wherein individuals vary in generating multiple, modifiable hypotheses to explain a social scenario (Hayward et al., 2018).

The idea of selecting from multiple mental state inferences has also long been present in theories of mindreading, such as the ToMM-SP theory (Leslie et al., 2004) and the Bayesian Theory of Mind (BToM) model (Baker et al., 2017). The ToMM-SP theory proposes that inhibitory selection, which involves a “Theory of Mind Module” that generates possible belief contents and a “Selection Processor” that selects among the generated beliefs, is the mechanism of making mental state attributions (Leslie et al., 2004). The BToM model similarly suggests that confidence levels or probabilities are initially assigned to candidate contents or hypotheses for mental state attributions, and then adjusted in specific circumstances, which then act as the criterion for selection of one’s final mental state attribution (Baker et al., 2017). To summarise, these models agree that while multiple hypotheses can be generated, some are rejected in the process of selecting a most appropriate interpretation, suggesting that the generation and selection of mental state interpretations can be studied as distinct processes. However, these models focus on mental state concepts such as beliefs (ToMM-SP model) and desires (BToM model) in simple, highly-constrained mindreading scenarios. The differences between generating and selecting appropriate mental state interpretations in more naturalistic, contextualised settings remain unexplored in the existing literature.

Understanding the distinction between generation and selection processes in mindreading is highly relevant to understanding how individuals navigate daily social activities across diverse contexts. The studies in the previous chapters have demonstrated the existence of multiple plausible interpretations of the same social scenario (Chapter 3), and that the perceived plausibility of these interpretations can be influenced by imposing contexts that favour certain interpretations over others (Chapter 5). These findings suggest that in real-life social scenarios, individuals often select among possibilities by taking context into consideration instead of simply “decoding” mental states from observable expressions of others. However, in Chapter 5 options were presented for participants to choose from, such that participants were not required to generate plausible interpretations on their own. Hence, the previous studies have not provided insight into whether context limits the generation of candidate interpretations. In other words, it remains unclear whether individuals take context into consideration when they are generating plausible candidate interpretations of others' mental states or only after that, when they select the most likely interpretation from the generated set.

6.1.2 Current study

The current study provides a first attempt to investigate the role of context in the generation of candidate interpretations separately from the selection of the best candidate, using a paradigm and method of modelling from decision making research by Morris et al. (2021). In their first experiment, Morris et al. (2021) manipulated the context (in their case, devaluing typically high-value food items in the given context of having just had a dental surgery) to increase participants' tendency to endorse items typically deemed unusual while making it less likely for participants to endorse an item deemed desirable (e.g., participants' favourite foods that required excessive chewing). The authors investigated whether the general desirability of the items and their desirability in the specific context predicted how likely the item would be (1) generated as a candidate option and (2) selected as the most desirable option in the given context. The authors found that the generation of candidate items was based on generalised evaluations of the items from past experience, whereas context-specific evaluations played a less important role. However, the selection of the most desirable option was based on the specific context. Although these effects do not logically entail that the same effects should be observed for mindreading, as the authors' investigation concerned the values of the items to participants, the paradigm can be adapted to investigate the generation and selection of mental state interpretations across contexts in mindreading.

In the current study, this paradigm was adapted to examine the effect of context on individuals' generation and selection of candidate mental state interpretations in ambiguous social scenarios. Participants were first required to generate candidate mindreading interpretations of social stimuli in a given context (“Context 1”) (e.g., the couple is sharing a meal on their anniversary), then select the best among those interpretations for that context and rate the likelihood of each of the generated interpretations in Context 1. Next, participants were required to rate the likelihood of each generated interpretation from Context 1 in an alternative context (“Context 2”) (e.g., the two colleagues are having dinner together after a work meeting). This approach enabled the investigation of whether the perceived likelihood of an interpretation of what a person is thinking or feeling in either context explained how likely it was to be generated as a plausible candidate in Context 1, and how likely it would be selected as the most likely interpretation in Context 1. Thus, it provided a novel way to address the key research question of the study: does context constrain both the generation of candidate interpretations and selection of the best interpretation?

If the process of generating candidate interpretations is constrained by context, then the candidate interpretations generated in Context 1 should be context-specific, hence rated likely for Context 1 but not an alternative context (Context 2). Additionally, the probability of generating an interpretation in Context 1 should not be explained by the interpretation’s perceived likelihood in Context 2. However, the alternative possibility is that candidate interpretations are generated with limited regard to the specific given context (see Morris et al., 2021), so interpretations generally considered likely for other contexts should also be generated. In this case, candidate interpretations considered likely for an irrelevant context (i.e., Context 2) are still expected to be generated in Context 1 more frequently compared with those considered unlikely for Context 2, despite their perceived likelihood for Context 1.

Nevertheless, as Studies 5-7 (Chapter 5) have shown the effect of context on selection of interpretations, it is predicted in the current study that while (1) context does not restrict the generation of candidate interpretations to those that fit only the current context, (2) context constrains what interpretation is selected as the best candidate.

6.2 Method

The current study was pre-registered on OSF prior to data collection (https://osf.io/3c5e9/). The study was approved by the Ethical Review Committee at the University of Birmingham.

6.2.1 Participants

An a priori power analysis was conducted with R 4.1.1 (R Core Team, 2021) and the simr package (Green & Macleod, 2016), which indicated 89% power (95% CI = [81.17%, 94.38%]) for detecting an assumed small effect size of 0.2 (Chen et al., 2010) for a fixed effect in a linear mixed model involving random intercepts and random slopes with a sample size of 300. A small effect size was specified due to the lack of existing data on the generation process of mindreading. Specifying a small effect size and recruiting a larger sample ensured that small but meaningful effects were not missed, as even a small effect could indicate whether a given context entirely constrains the generation of candidate interpretations to those fitting the current context. The following screening criteria were specified in the recruitment of 300 participants via Prolific: participants had to be aged between 18 and 25, were UK residents, spoke English as their first language, had not been diagnosed with ASD, and had not participated in studies 1-7. The screening criteria were imposed to match the sample with that of Studies 5 to 7, in which the contexts were shown to influence mental state interpretations. Three participants did not fulfil the age criterion: two participants were aged 26, while one participant was 43. The two participants aged 26 were kept in the dataset as their age was still very close to the upper limit (25), but the participant aged 43 was screened out from the analysis. Another participant was excluded as they provided the same rating response in 100% of the rating questions. Hence, the final sample included 298 individuals aged between 18 and 26 (148 female, M_{age} = 22.58).

Among the participants, 77.9% (n=232) were monolingual, while 16.1% (n=48) were bilingual and 6.0% (n=18) spoke more than two languages. Just over half (56.7%) held a Bachelor's degree or above. Participants identified their ethnicity according to the descriptions recommended by the United Kingdom Office for National Statistics (ONS). Most of the participants were White (70.5%, n=210), 16.8% (n=50) were Asian, 6.0% (n=18) were mixed, 0.3% (n=1) was Black and 0.3% (n=1) identified as an Other ethnic group.

6.2.2 Study design and procedure

Informed written consent was obtained before all participants participated in this study. All participants completed an online questionnaire on Qualtrics. After the participants read the instructions, five pictures each depicting a social scenario were shown.

The five pictures were presented twice, once in each of the two main blocks of the questionnaire. Figure 6.1 illustrates the flow of the two blocks.

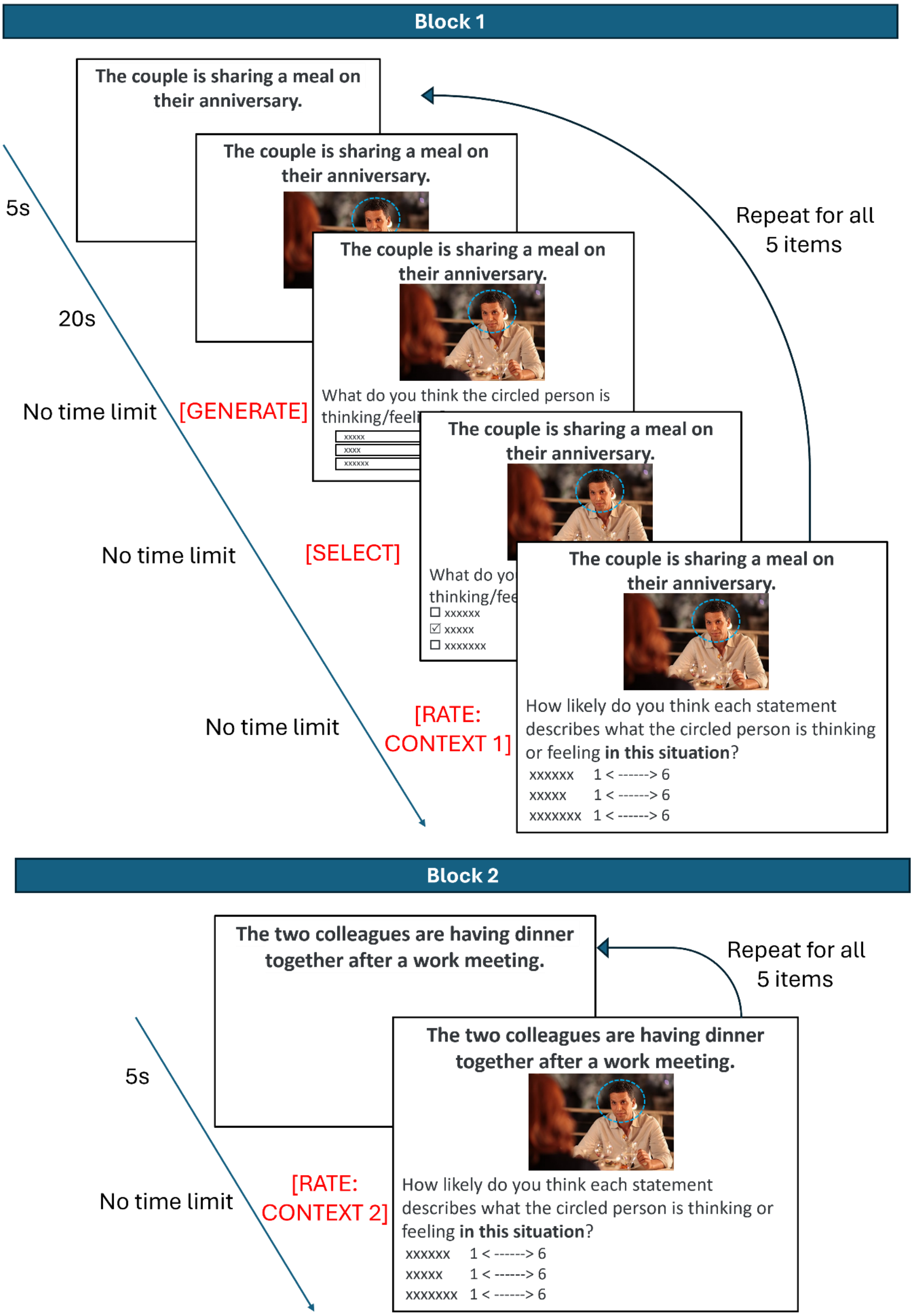

Block 1
The couple is sharing a meal on their anniversary.
5s
20s
No time limit
[GENERATE]
What do you think the circled person is thinking/feeli
xxxxx
xxxx
xxxxxx
No time limit
[SELECT]
What do yo
thinking/fee
☐ xxxxxx
☑ xxxxx
☐ xxxxxxx
No time limit
[RATE: CONTEXT 1]
How likely do you think each statement describes what the circled person is thinking or feeling in this situation?
xxxxxx 1 < ------> 6
xxxxx 1 < ------> 6
xxxxxxx 1 < ------> 6
Repeat for all 5 items
Block 2
The two colleagues are having dinner together after a work meeting.
5s
No time limit
[RATE: CONTEXT 2]
How likely do you think each statement describes what the circled person is thinking or feeling in this situation?
xxxxxx 1 < ------> 6
xxxxx 1 < ------> 6
xxxxxxx 1 < ------> 6
Repeat for all 5 items

Figure 6.1. The figure illustrates the presentation of items and questions in the two experimental blocks. The interpretations rated in Block 2 were generated by in Block 1.

In the first block, before each picture was presented, participants read a sentence describing the contextual background of the picture (Context 1). The sentence described the relationship between the characters or the general background to the interaction (e.g., “The couple is sharing a meal on their anniversary”). The picture was then presented after 5s, and participants were given 20s to think about possible interpretations of what a target character in the picture was thinking or feeling. Participants were then required to enter all the possible interpretations they had thought of during the 20s, up to a maximum of 10 interpretations, even if they only thought of them briefly and soon rejected them. The next page then showed participants the responses they had just entered and instructed the participants to select only one entry as the most likely interpretation in the given context. After that, participants rated the likelihood of every interpretation they entered on a 6-point scale, ranging from 1 (extremely unlikely) to 6 (extremely likely).

Participants read instructions regarding the next block before the second block commenced. In the second block, the same pictures were presented in the same sequence as in the first block, but the contextual information was changed (Context 2; e.g., the two colleagues are having dinner together after a work meeting). Participants were shown the list of interpretations they had entered earlier in the first block to rate how likely they thought these interpretations described the picture when taking the new context into consideration.

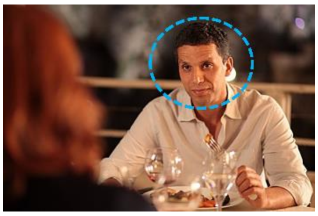
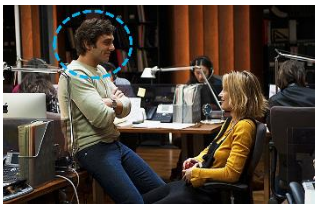
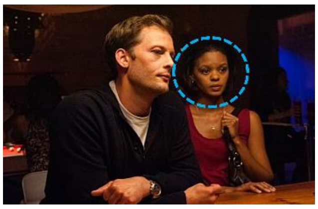
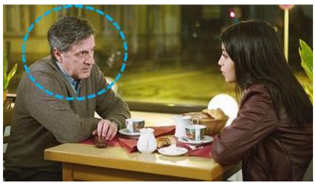
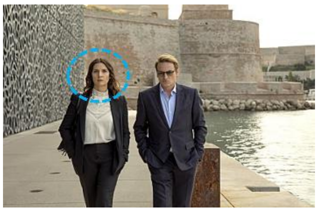
Context 1 (i.e., a low-frequency context) always presented a context designed to prompt an infrequent interpretation of the item when the item was presented context-free, based on Studies 5d-7. Context 2 (i.e., the high-frequency context) always presented a context that prompted a frequent interpretation of the item when it was presented context-free (see Studies 5d-7). The presentation order of pictures in both blocks was counterbalanced between participants using a Latin square design, following five unique sequences. The presentation order of the interpretations was randomised each time participants had to rate the likelihood of each one.

The session took around 15 minutes and participants received £2.25 after completing the study.

6.2.3 Materials

The same five pictures, as well as the low-frequency (Context 1) and high-frequency context (Context 2) descriptions from Study 7 were used in this study. The pictures and contexts presented are summarised below in Table 6.1.

Table 6.1. List of pictures and contexts presented in Study 8.

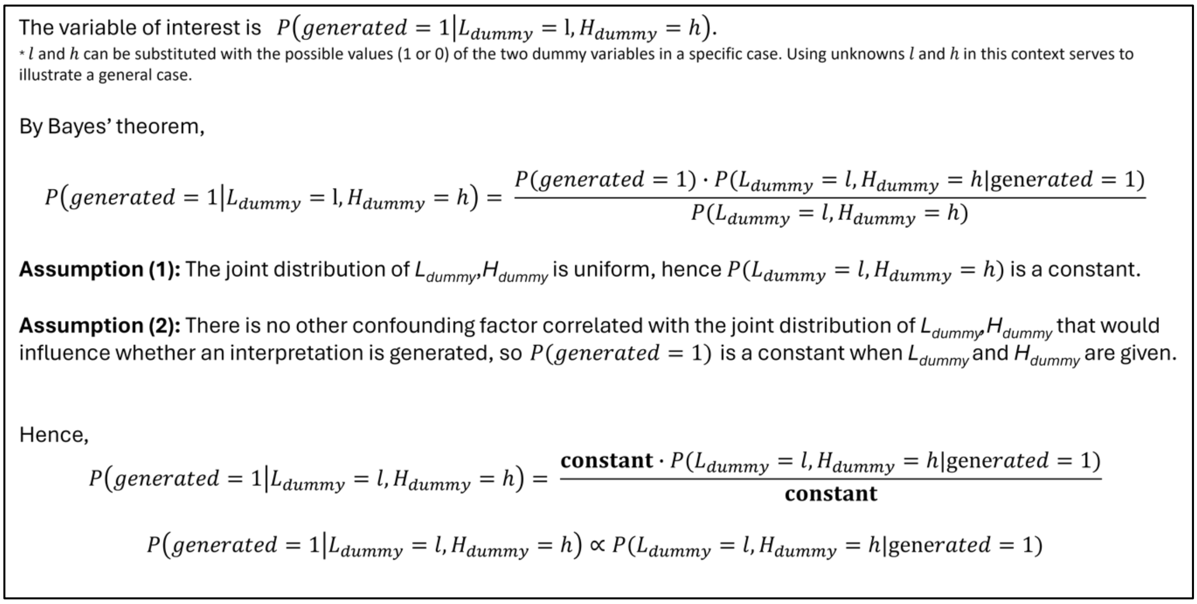
Picture	Context 1	Context 2
P3 	The couple is sharing a meal on their anniversary.	The two colleagues are having dinner together after a work meeting.
P5 	He just got promoted to leader of their team.	The colleagues have found that they are both free after work today.
P8 	They've been told that their reservation was cancelled.	They've been told that it hasn't yet been possible to contact their daughter.
P9 	He just invited his daughter out to tell her his decision to divorce her mother.	They are meeting because his daughter called him.
P10 	The couple is on their way home from lunch.	They are on the way to handling a difficult assignment.

6.2.4 Analyses and data pre-processing

Generation of candidate interpretations. In the analysis of the effect on generation of candidate interpretations, the target dependent variable was the probability of generating a candidate interpretation in Context 1 given its perceived likelihood in Context 1 and in Context 2. This probability is unmeasurable: to calculate the probability of generating the interpretations in Context 1, the number of interpretations generated should be divided by the number of interpretations that can be possibly generated (an unknown value). However, a proxy of the probability of interest was calculated following the method used in Morris et al. (2021), in which the analysis was made possible by recoding data and imposing specific assumptions as follows. For convenience, ratings for Context 1 are labelled L (for low frequency context) and the rating for Context 2 are labelled H (for high frequency context) in the following explanation of the analyses.

The perceived likelihood ratings in both contexts were recoded into a binary variable based on whether the rating was higher than (1) or lower than midpoint of the scale (0), to allow for analysing the data with simplifying assumptions, as will be explained below. This created four possible combinations of the values of the two dummy variables across contexts: (1) L_{dummy}=1 & H_{dummy}=1, (2) L_{dummy}=0 & H_{dummy}=1, (3) L_{dummy}=1 & H_{dummy}=0, and (4) L_{dummy}=0 & H_{dummy}=0.

There were two assumptions related to the joint distribution of L_{dummy},H_{dummy}. First, the joint distribution was assumed to be uniform. In other words, it was assumed that there were equal numbers of possible interpretations that fit any of the four combinations. The second assumption was that the joint distribution of L_{dummy},H_{dummy} did not correlate with any confounding factor that would influence whether an interpretation would be generated, that is, the perceived likelihood of an interpretation in the two contexts was the only factor that influenced whether an interpretation would be generated. With these two assumptions and using Bayes' theorem, the probability of generating an interpretation given a specific combination of rating values (i.e., the variable of interest) is directly proportional to the proportion of responses that match the corresponding combination of values of the two dummy variables, as illustrated in Figure 6.2. Hence, this latter proportion, which can be calculated from the data collected (as explained in Appendix D), can be used as a proxy of the probability of generating an interpretation in Context 1. This proxy was, therefore, used as the dependent variable in the linear mixed model analysis.



The variable of interest is $P(generated = 1|L_{dummy} = \mathrm{l}, H_{dummy} = h)$.

* l and h can be substituted with the possible values (1 or 0) of the two dummy variables in a specific case. Using unknowns l and h in this context serves to illustrate a general case.

By Bayes' theorem,

$$P(generated = 1|L_{dummy} = \mathrm{l}, H_{dummy} = h) = \frac{P(generated = 1) \cdot P(L_{dummy} = l, H_{dummy} = h|\text{gener}ated = 1)}{P(L_{dummy} = l, H_{dummy} = h)}$$

Assumption (1): The joint distribution of L_{dummy},H_{dummy} is uniform, hence $P(L_{dummy} = l, H_{dummy} = h)$ is a constant.

Assumption (2): There is no other confounding factor correlated with the joint distribution of L_{dummy},H_{dummy} that would influence whether an interpretation is generated, so $P(generated = 1)$ is a constant when L_{dummy} and H_{dummy} are given.

Hence,

$$P(generated = 1|L_{dummy} = l, H_{dummy} = h) = \frac{\textbf{constant} \cdot P(L_{dummy} = l, H_{dummy} = h|\text{gener}ated = 1)}{\textbf{constant}}$$

$$P(generated = 1|L_{dummy} = l, H_{dummy} = h) \propto P(L_{dummy} = l, H_{dummy} = h|\text{gener}ated = 1)$$

Figure 6.2. For each item, the proportion of a participant's interpretations matching each combination of the two dummy variables' values ($P(L_{dummy} = l,\ H_{dummy} = h|generated = 1)$) is directly proportional to the probability of generating an interpretation that fits the corresponding combination of perceived likelihood in the two contexts ($P(generated = 1|L_{dummy} = l,\ H_{dummy} = h)$) for each item, and hence, can be used as a proxy of the latter. In sum, the dependent variable was a vector of four numbers corresponding to the proportion of responses falling into the four possible combinations of high versus low likelihood ratings for the two contexts, for each item for each participant (see Appendix D).

In the linear mixed model, the fixed effects of the two dummy variables were the key to addressing the research question. If context does not fully constrain the generation of candidate interpretations, the fixed effects of both Context 1 and Context 2 should be significant in explaining the probability of generating an interpretation in Context 1. The analysis was conducted using the packages lme4 (Bates et al., 2015) and lmerTest (Kuznetsova et al., 2017) in R 4.1.1 with the REML estimator. As specified in the preregistration, a full random structure was first specified. If the model failed to converge, a model without random slopes would be specified.

Selection of the best interpretation. To address the question of selecting the most likely interpretation in Context 1, a logistic mixed model was specified. The dependent variable was binary: it was coded 1 if the interpretation was selected, or 0 if not. For each participant, only one interpretation would be assigned a value of 1 for each item. The predictors were the raw likelihood ratings of interpretations in Context 1 and Context 2.

If context constrains the selection of the best interpretation, it was expected that the fixed effect of only Context 1 but not Context 2 should be significant. The logistic mixed model analysis was conducted using the packages lme4 and lmerTest in R 4.1.1 with the ML estimator. The model specification started with a full random structure. Random slopes would be removed if the model failed to converge or had a singular fit.

For both models addressing generation and selection, any further problems with model convergence or model fit would be handled by specifying a Bayesian model with the full random structure using the package brms (Bürkner, 2018) as an alternative, with reference to research showing the adoption of Bayesian model could practically solve convergence issues (Kimball et al., 2019).

6.3 Results

6.3.1 Generation of candidate interpretations

Descriptive statistics. Table 6.1 summarises the proportion of all interpretations generated by all participants in Context 1 that were then rated as likely (likelihood rating above scale mid-point) and unlikely (below scale mid-point) in Context 1 (the low frequency context) and Context 2 (the high frequency context). The vast majority (81.7%) of the interpretations generated in Context 1 were rated as likely in Context 1, while only slightly over half (52.5%) were rated as likely in the other context (Context 2). This observation suggests that overall, individuals were slightly more likely to generate interpretations that also fit an alternative Context 2 than those that did not. However, only 6.18% of the interpretations were rated as likely in the alternative context (Context 2) but unlikely in the context where the interpretations were generated (Context 1). This suggests that it was unlikely for individuals to come up with candidate interpretations that suited an alternative context more than the current given context.

Table 6.1. Proportion of interpretations categorised as likely and not likely across contexts.

		Context 2		**Subtotal**
		Likely	Not likely	
Context 1	Likely	46.3%	35.4%	81.7%
	Not likely	6.18%	12.1%	18.3%
Subtotal		52.5%	47.5%	

As the data summarised above were nested within participants, mixed models were adopted for formally testing the hypotheses.

Mixed model results. If the generation of candidate interpretations was not constrained by context, then the probability of a candidate interpretation being generated in Context 1 should be associated with perceived likelihood in both Context 1 and Context 2. The linear mixed model was first specified to include a random intercept and the random effects of both dummy variables (perceived likelihood in Context 1 and that in context 2). The model had a singular fit, so random slopes were removed to reduce model complexity. However, the simplified model still resulted in a singular fit, probably due to extremely low between-participant variance (estimate = .00, *SD* = .00).

As an alternative, a Bayesian linear mixed model that included a random intercept and random effects of both predictors was specified using the brms package in R 4.1.1 with default, weakly informative priors set by the package, due to a lack of strong prior expectations regarding the magnitude of the effects (Bürkner, 2017). The zero/one inflated Beta model family was adopted to cater for the nature of the dependent variable, which was a proportion that varied between 0 and 1 (0 and 1 inclusive) (Liu & Eugenio, 2016).

Results showed that the likelihood rating in both Context 1, estimate = 0.65, *SE* = 0.03, 95% credible interval = [0.58, 0.71], and Context 2, estimate = .10, *SE* = .04, 95% credible interval = [.02, .16], explained the probability of an interpretation being generated in Context 1. In other words, when an interpretation was considered likely in Context 1, the probability of it being generated was increased by an estimated value of 65% compared to an unlikely interpretation in Context 1, controlling for the likelihood of the interpretations in another context (Context 2). Likewise, when an interpretation was considered likely in another context (Context 2), the probability of it being generated was increased by an estimate of 10% compared to an unlikely interpretation in Context 2, controlling for the likelihood of the interpretations in the given context (Context 1). The 95% credible interval of the two estimates did not overlap, indicating that the probability of generating an interpretation was explained to a larger extent by its perceived likelihood in Context 1 than that in Context 2. Table 6.2 summarises the output of the model.

Table 6.2. Output table of the Bayesian model (zero/one inflated Beta model family) for explaining generation of candidate interpretations in Context 1.

	Generation		
Predictors	*Estimates*	*std. Error*	*CI (95%)*
Intercept	-0.87	0.03	[-0.93 – -0.80]
Context 1 (dummy)	0.65	0.03	[0.58 – 0.71]
Context 2 (dummy)	0.10	0.04	[0.02 – 0.16]
phi	7.27	0.21	[6.88 – 7.68]
Random Effects			
σ^2	1.00		
τ_{00} PID	0.06		
τ_{11} PID. Context1(dummy)	0.02		
τ_{11} PID. Context2(dummy)	0.15		
ρ_{01}			
ρ_{01}			
ICC	0.04		
N PID	298		
Observations	5960		
Marginal R^2 / Conditional R^2	0.016 / 0.023		

To check the robustness of the output, the Bayesian mixed model was rerun by replacing the Beta distribution with the Gaussian distribution in the specification of model family. As the Gaussian distribution was more general and less tailored to the specific distribution of the current data, it was used to examine the robustness of the findings across models with differing underlying assumptions. The default weakly informed priors were adopted. Both fixed effects were replicated, showing that both likelihood ratings in Context 1 (estimate = 0.33, *SE* = 0.01, 95% credible interval = [0.31, 0.34]) and Context 2 (estimate = 0.03, *SE* = 0.01, 95% credible interval = [0.01, 0.05]) explained the generation of an interpretation in Context 1. Table 6.3 summarises the model output.

Table 6.3. Output table of the Bayesian model (Gaussian model family) for explaining generation of candidate interpretations in Context 1.

	Generation		
Predictors	*Estimates*	*std. Error*	*CI (95%)*
Intercept	0.07	0.01	[0.06 – 0.09]
Context 1 (dummy)	0.33	0.01	[0.31 – 0.34]
Context 2 (dummy)	0.03	0.01	[0.01 – 0.05]
Random Effects			
σ^2	0.06		
τ_{00} PID	0.01		
τ_{11} PID. L_dummy	0.01		
τ_{11} PID. H_dummy	0.02		
ρ_{01}			
ρ_{01}			
ICC	0.10		
N PID	298		
Observations	5960		
Marginal R^2 / Conditional R^2	0.301 / 0.374		

6.3.2 Selection of the most likely interpretation

Descriptive statistics. Table 8.4 shows the descriptive statistics of the likelihood ratings for the interpretations that were selected and not selected as the best interpretations in Context 1, across items.

Table 6.4. Descriptive statistics of likelihood ratings for interpretations selected and not selected as the best interpretation in Context 1.

	Likelihood ratings in Context 1 (range = 1-6)		**Likelihood ratings in Context 2 (range = 1-6)**	
	Mean	*SD*	*Mean*	*SD*
Selected	5.48	0.71	3.76	1.76
Not selected	4.34	1.18	3.46	1.76

Mixed model results. If selection of the most likely interpretation in a scenario depends on the context, then the perceived likelihood rating of a candidate in Context 1 but not Context 2 should explain its possibility of being selected as the best interpretation in Context 1. To test this hypothesis, the first mixed model was specified with both a random intercept and random slopes, but the model failed to converge. Hence, a simplified model including only the random intercept but not the random slopes was specified. Results showed that the selection of the best interpretation in Context 1 was only predicted by its perceived likelihood in Context 1 with a log-odds of 1.44 (z = 26.50, p < .001) corresponding to an odds ratio of 4.20, but not its perceived likelihood in Context 2. The output of the analysis is summarised in Table 6.5.

Table 6.5. Output table of the logistic mixed model predicting selection of the best interpretation in Context 1.

	Selection					
Predictors	*Log-Odds*	*std. Error*	*CI*	*Statistic*	*Odds ratio*	*p*
(Intercept)	-8.31	0.29	[-7.34 – -6.42]	-28.78	0.00	<.001***
Context 1 (rating)	1.43	0.05	[1.33 – 1.54]	26.44	4.19	<.001***
Context 2 (rating)	-0.02	0.02	[-0.07 – 0.02]	-0.82	0.98	.286
Random Effects						
σ^2		3.29				
τ_{00} PID		0.25				
ICC		0.07				
N PID		298				
Observations		6551				
Marginal R^2 / Conditional R^2		0.451 / 0.489				

*** *p* < .001

To check the robustness of the effects, the complex model with full random components was rerun with the Bayesian approach using the brms package and package default priors. Results from the simplified model were replicated, showing that only perceived likelihood in Context 1 but not Context 2 explained whether an interpretation was selected as the best interpretation in Context 1, with a log-odds of 1.55 (95% credible interval = [1.42, 1.69]) corresponding to an odds ratio of 4.70, as summarised in Table 6.6.

Table 6.6. Output of the Bayesian model for predicting selection of the best interpretation in Context 1.

	Selection			
Predictors	*Log-Odds*	*std. Error*	*CI (95%)*	*Odds Ratio*
Intercept	-8.97	0.36	[-9.70 – -8.27]	0.00
Context 1 (rating)	1.55	0.07	[1.42 – 1.68]	4.69
Context 2 (rating)	-0.02	0.02	[-0.06 – 0.02]	0.98
Random Effects				
σ^2	3.29			
$\tau_{00\ \mathrm{PID}}$	7.42			
$\tau_{11\ \mathrm{PID.Context1}}$	0.21			
$\tau_{11\ \mathrm{PID.Context2}}$	0.00			
ρ_{01}				
ρ_{01}				
ICC	0.20			
N_{PID}	298			
Observations	6551			
Marginal R^2 / Conditional R^2	0.238 / 0.240			

6.3.3 Additional analysis: inter-item correlation in the number of candidates generated

An additional analysis was conducted to examine participants' consistency in the number of candidate interpretations they generated across cross items. This analysis was conducted as an extension to Study 3 (Chapter 3) to provide further insight into using propensity to generate multiple interpretations as a potential index of consistent individual differences in mindreading. The average number of interpretations generated for each item varied from 4.21 to 4.58 across participants. Echoing the findings from Study 3 (Chapter 3), the inter-item correlations were high concerning the number of interpretations generated, as shown in Table 6.7.

Table 6.7. Zero-order correlation (Spearman correlation) matrix of number of candidate interpretations generated across items.

	P3	**P5**	**P8**	**P9**
P5	.58***			
P8	.62***	.60***		
P9	.60***	.65***	.58***	
P10	.54***	.57***	.55***	.54***

*** *p* < .001

6.4 Discussion

6.4.1 Summary and implications

This study aimed to investigate if context constrained selection and generation of candidate mental state interpretations in ambiguous social stimuli. It was predicted that context would not limit candidate interpretations generated to only those considered likely in the given context, but would strongly limit the selection of the best candidate. As hypothesised, the generation of candidate interpretations was not entirely shielded from the constraints of the given contexts, but contexts constrained the selection of the best pick among the generated. In other words, selection was highly context-dependent but generation was not. Additionally, this study has provided support for the previous findings (Study 3) showing consistent individual differences in generating candidate interpretations.

The current findings reiterate the importance of context in the process of selecting a good interpretation in mindreading “decoding” tasks in which participants interpret mental states based on observable cues (Meinhardt-Injac et al., 2020), most likely, facial expressions (e.g., Cambridge mindreading face battery, Baron-Cohen & Hill, 2006; Faces test, Adoplhs et al., 2002; Baron-Cohen et al., 1997; Reading the Mind in the Eyes Task, Baron-Cohen et al., 2001). This echoes the findings from Studies 5-7 (Chapter 5).

Furthermore, the current findings provide support for studying generation and selection as separate processes involved in mindreading. Measurements of mindreading in the existing literature predominantly consider the appropriateness of a response, which focus on the selection process, or conflate generation and selection. Although other research has distinguished the appropriateness of mindreading responses and one's propensity to mindread (e.g., Devine & Apperly, 2022; Carpenter et al., 2016), the behavioural assessment of the latter usually rests on whether a mindreading interpretation is spontaneously produced or the proportion of mental state descriptions in open-ended responses. The tendency to generate multiple interpretations for the same target person has been seldom considered in the literature with one notable exception (i.e., Hayward et al., 2018), in which the authors presented comic strips to children aged between 7 and 17, who were prompted repeatedly to provide explanations for a character's behaviour until they could not think of any more explanations. Hence, the stand-alone process of generating candidate interpretations seems to be understudied in the existing literature.

The idea of studying generation and selection corroborates existing theoretical models of mindreading but also provides unique contribution to the study in the field of mindreading. While the ToMM-SP (Leslie et al., 2004) has not specified how the Selection Processor makes decisions on selecting the most appropriate mental state, the current findings show that the selection process involves the consideration of context. The current findings also suggest that these two distinct processes are not only involved in reasoning about an agent's beliefs and desires in highly-constrained settings as shown in studies of the BToM model (e.g., Baker et al., 2011), but also more general thoughts and feelings in naturalistic social scenarios. To illustrate with the study by Baket er al. (2011), a sample stimulus presented involved a 2-D spatial configuration featuring a cartoon agent, multiple food trucks and objects obstructing the agent's view. The view the agent was able to see was manipulated to control the agent's beliefs about the location of the food trucks, while the agent's desires were limited to their preference for a particular food truck or another. This setting was highly specific, unlike the stimuli in the current study.

6.4.2 Limitations and future research directions

One limitation of the current study was the inability to examine whether the effect of context on the generation and selection of mental state interpretations is time-sensitive. In the current study design, following Morris et al. (2021), participants were asked to generate possible interpretations within a brief period (20 seconds in the current study) and then recall them, with no restrictions on the order of recall. Therefore, the current study did not explore whether interpretations generated earlier were less influenced by context than those generated later. This issue of time-sensitivity could be addressed by asking participants to recall their thoughts in sequence. Alternatively, investigating the impact of speeded responses or increased cognitive load on considering possible interpretations might shed light on the role of time and cognitive effort in generating and selecting interpretations in a given context.

A notable observation from our results was the relatively small effect size of likelihood ratings in Context 2 compared to Context 1 in predicting the generation of candidate interpretations in Context 1. This contrasts with the findings from Morris et al. (2021), where they observed a larger effect of generalised evaluations than context-specific evaluations on the generation of candidate items. While a definitive explanation for this finding is lacking, it is possible that interpretations fitting Context 2 were not as general as in Morris et al.'s (2001) study, leading to lower likelihood of their generation by default, assuming context would not have constrained generation of candidate interpretations. Future research could further investigate the effect of context by manipulating the similarity between the two contexts and examining whether a more similar or dissimilar, irrelevant context influences how likely an interpretation fitting the irrelevant context is generated in the initially given context.

6.4.3 Conclusion

Study 8 aimed to explore the role of context in both generation and selection of mental state interpretations of ambiguous social stimuli. Results showed that context strongly affected selection while its effect on generation was weaker. The findings emphasised the importance of context in mental state interpretations and supported the distinction between generation and selection processes in mindreading.

Chapter 7

General discussion

7.1 Summary and synthesis

7.1.1 Overarching aims and summary

This thesis has addressed research gaps in the study of individual differences in adult mindreading by focusing on measurement challenges, the conceptualisation of mindreading success, and examining the processes of generation and selection in context.

First, to address the question on measurement challenges, Chapter 2 presented a systematic review evaluating the sensitivity of existing mindreading measures to performance variance in adults and their psychometric properties. To identify reliable indices of individual differences in adult mindreading, Chapters 3 to 5 investigated inter-item correlations of participants' alignment within groups. Furthermore, Chapter 5 specifically addressed participants' flexibility in adjusting interpretations based on varying contexts.

Second, to investigate the conceptualisation of mindreading success, Chapter 3 discussed how mindreading success could be defined in mental state tasks where the ground truth of targets' mental states is not directly accessible. Alignment was proposed as an alternative to accuracy, and the chapter introduced a novel method to quantify alignment from open-ended data. Chapters 4 and 5 further challenged the notion of accuracy by investigating changes in alignment by manipulating the process that participants are required to engage in due to varied format, and context provided in the task, respectively.

Third, the generation and selection processes in mindreading were considered throughout this thesis. In Chapter 3, where a correlational analysis was conducted to test the association between participants' tendency to generate unique interpretations and their alignment within groups. Chapter 6 then examined the role of context in the generation and selection of interpretations as distinct processes, using a novel paradigm and a modelling approach.

This final chapter will first summarise the chapter aims and findings, then integrate the findings from all chapters to answer the overarching themes of the current thesis, followed by discussing overarching implications, and finally discuss general limitations of the current studies and future research directions, ending with a concluding remark.

7.1.2 Summary of chapter aims and findings

Chapter 2. Chapter 2 presented a systematic review identifying 75 existing measures used to assess individual differences in mindreading in neurotypical adults. This review examined how these measures were administered and critically evaluated their measurement characteristics, focusing on sensitivity to individual differences in mindreading performance, reliability, validity, and the interrelations among tasks. The review revealed inconsistencies in the administration of measures across studies, with variations in stimuli and response formats even for notionally identical tasks. Among open-ended measures, the most tasks focused on accuracy rather than propensity in mindreading.

For measures where it was possible to examine the average percentage of maximum possible score (POMP), approximately half exhibited ceiling effects (i.e., average scores were >85% of the maximum possible score). There was a notable lack of evidence regarding reliability and validity as this information were often not reported. The analysis thus concentrated on the top eight most commonly used measures. Although the reliability of these measures was generally satisfactory, information on test-retest reliability was limited. Validity was primarily assessed in terms of convergent validity and known-group validity, with seven out of the eight tasks showing satisfactory results in these areas. However, evidence for discriminant validity was limited, despite being mostly positive, and there was even less evidence for criterion-related validity, indicating that real-life outcomes of mindreading in neurotypical adults may be understudied. Moreover, despite its popularity, the Reading the Mind in the Eyes Task (RMET; Baron-Cohen et al., 2001) did not exhibit the best psychometric properties. Furthermore, inconsistent evidence was found regarding intercorrelations among tasks, with some measures not correlating with any others, even after correcting for attenuation.

These findings underscore the need for more empirical research on the psychometric properties of existing measures and the development of measures with better psychometric characteristics. Future research adopting a latent variable approach to determine whether mindreading can be conceptualised as a common latent factor represented by multiple sub-abilities that are interrelated is needed.

The systematic review provided a basis for exploring alternatives to the accuracy criterion and examining how task format and context influence mental state interpretations in subsequent chapters.

Chapter 3. Chapter 3 presented a series of three studies that aimed to address the assumption that a “correct” answer determined by a small group of experimenters, panel members, or pilot participants served as a proxy for a single, invariant ground truth about a target’s mental states. This chapter explored alignment as an alternative to accuracy in characterising mindreading success with a novel method to operationalise alignment. The chapter also evaluated the tendency to generate multiple unique mental state interpretations and alignment as possible indices of individual differences in mindreading.

Photos featuring individuals of varied ages, genders, and ethnicities engaging in naturalistic social interaction scenarios were used as stimuli. This approach differed from other existing tasks that presented either completely decontextualized photos of faces or the eye region (e.g., Adoplhs et al., 2002; Baron-Cohen et al., 1997; 2001) or video clips of social interactions. The characteristics of the current stimuli, being both naturalistic in the social scenarios portrayed and minimally contextualised (though not entirely decontextualised), allowed for the investigation of a wide range of plausible mental state interpretations and the manipulation of context in later chapters (Chapters 5 and 6).

Study 1 provided pilot data for Studies 2 and 3. Together, Studies 1 to 3 found evidence for differences in how older and younger participants interpreted the mental states of the same targets. The results challenged the assumption that the best description held for individuals across different groups. Study 3 further found evidence for multiple interpretations for each item, challenging the assumption of a single proxy to ground truth when actual ground truth was unknown. However, the tendency to generate multiple unique mental state interpretations was found to be a potential candidate for reliably indexing individual differences in mindreading, even when controlling for verbosity. Alignment scores were not found to be interrelated across items, suggesting that alignment of open-ended interpretations of ambiguous picture stimuli might not provide a reliable indicator of individual differences in mindreading. An additional exploratory analysis was conducted to examine if the propensity to generate multiple interpretations correlated with alignment scores, but the correlations were weak and not statistically significant. This corroborated research suggesting propensity as an independent facet of mindreading from accuracy, with the former being a motivation-related component while the latter involves criteria to determine the appropriateness of an interpretation.

It was speculated that the lack of inter-item correlation on alignment scores was due to the open-ended format of the task and the absence of given background information, which permitted a broad range of possible interpretations. These factors were then examined in Chapter 5 by investigating the inter-item correlations of alignment scores when the task was changed to a forced-choice format and, further, by introducing context information to constrain interpretations. The task was first converted to a forced-choice format in Chapter 4.

Chapter 4. Chapter 4 compared the alignment of mental state interpretations in a newly recruited sample completing a forced-choice version of the task with that of participants completing the open-ended version of the task (i.e., the younger adult group from Study 3). In the forced-choice version, the four options were derived from popular mental state categories identified in Study 3. Results showed that participants' alignment within the same-format sample (forced-choice) was significantly higher than alignment with the open-ended-format sample, indicating that the format of the task, which involved the process of recognition versus generation, influenced preferences for different mental state interpretations. Consistent with Chapter 2, this finding underscored the importance of consistency in task administration across studies, suggesting that the validity of a "correct" answer could be affected by changes in task format.

Chapter 5. Chapter 5 presented three sets of studies (Studies 5a-d to 7) focused on the role of context in influencing mental state interpretations. Studies 5a-d involved stimulus refinement and tested the effect of context with a between-participant design, which was then tested in a within-participant design in Study 6. Study 7 not only replicated the context effect in Study 6 but also operationalised flexibility as the tendency to adjust mental state interpretations to align with the majority, using a large sample of 140 participants. The inter-item correlation of flexibility across items was examined to determine its reliability as an index of individual differences in mindreading. Additionally, the inter-item correlations of participants' alignment scores calculated using the same method as in Study 4 (1) without context and (2) within each manipulated context were analysed to investigate whether the lack of inter-item correlation in Study 3 was due to the open-ended format or minimal contextual constraints.

Results from all three sets of studies consistently showed that context significantly influenced mental state interpretations. Specifically, a less-favoured interpretation in a decontextualised situation became more popular when a context favouring it was provided. This effect was observed in both between-group comparisons and within-participant adjustments, as participants changed their interpretations when presented with a different context. The observation of within-participant adjustment provided a basis for examining flexibility to adjust mental state interpretations with varying contexts. However, flexibility scores were not interrelated across items.

Additionally, alignment scores were calculated for participants when stimuli were presented with no context information, as well as with contexts suggesting low-frequency and high-frequency interpretations. None of these scenarios showed satisfactory inter-item correlations for alignment scores. The lack of inter-item correlations suggested that neither flexibility nor alignment scores were suitable for reliably indexing individual differences in adult mindreading, at least with the current task. The lack of inter-item correlations in alignment scores further indicated that the absence of inter-item correlation in Study 3 was not likely due to its open-ended format or minimal context constraints.

Chapter 6. Chapter 6 extended the investigation of the role of context in mental state generation, building on findings from Chapter 5. This chapter not only examined the selection of the most appropriate interpretation but also the generation of candidate interpretations within a given context. Adapting a novel paradigm adapted from value-based decision-making research that showed context constrains selection but not generation of plausible items (Morris et al., 2021), Study 8 tested whether context similarly constrained the selection and generation of mindreading interpretations. Participants were asked to generate plausible mental state interpretations in a given context, to rate their perceived likelihood within that context, and then to rate their perceived likelihood in a different context.

The results indicated that, although context influenced the candidate interpretations generated, the generation of these interpretations was not entirely constrained by the given context. In contrast, contexts strongly constrained the selection of the best pick among the generated mindreading interpretations, meaning that the selected best interpretation was highly context-dependent. Additionally, a further analysis examined participants' consistency in the number of candidate interpretations they generated across different items. The number of candidate interpretations generated showed strong correlations across items, supporting the finding from Study 3 that the tendency to generate multiple plausible mental state interpretations was a reliable indicator of individual difference in mindreading. This tendency might be more relevant to measuring propensity, as it did not involve judging the appropriateness of the generated interpretations.

7.1.3 Summary of answers to overarching themes

Measurement challenges. Chapter 2 revealed inconsistencies in task administration, ceiling effects in a considerable number of measures, and a limited evidence base for evaluating reliability and validity for existing measures. The tendency to generate multiple interpretations of the mental states of targets in ambiguous visual stimuli was a promising candidate for studying individual differences in mindreading. In contrast, alignment and flexibility of responses to the current task may not be reliable indices of mindreading. The lack of inter-item correlation for flexibility scores could be due to limited observed variance, but the reasons for the lack of alignment correlation remain unclear. It is unlikely that these issues were due to the open-ended nature of the task or the absence of context.

Defining mindreading success. Chapter 3 challenged the assumption that "accuracy" determined by a small group was an effective proxy to a single ground truth in mental state interpretations. Alignment was introduced as an alternative to accuracy, and results revealed significant differences in interpretation across age groups in addition to multiple popular interpretations for all items. Chapters 4 and 5 further challenged the notion of accuracy, which was assumed to be invariant in mental state “decoding”, by demonstrating that task format and context significantly influenced interpretations.

Generation and selection in mindreading. Chapters 3 and 6 addressed the distinction between generation and selection processes. Chapter 3 studied the two processes by, first, calculating tendencies in generating multiple interpretations separately from the alignment within group, and second, correlating propensity with alignment, showing weak and insignificant correlations between the two. Chapter 6 explored whether the role of context differed between the two processes, finding that selection of the most appropriate mental state interpretation was largely dependent on context, while the generation of candidate interpretations was influenced by context, but to a lesser extent.

7.2 Novelty of current research and implications

7.2.1 Novelty of current research

The studies reported in this thesis were built on existing empirical and theoretical research but contribute to the field in three ways.

First, the scoring of participants was not based on a set of pre-determined “correct” answers but was achieved through crowdsourcing. In other words, participants’ scores depended on their agreement with fellow participants. This method departed from the conventional definition of mindreading success based on “accuracy”, which was unknown in the current case, to “alignment” as an alternative definition of mindreading success. Particularly, by calculating alignment scores in a way that took the proportion of agreement with others into consideration, the analysis accounted for the full set of variation in responses to detect differences not just in the majority choice between people, but also quantitative differences in their mental state interpretations. This was especially important given that Chapter 3 found more than one popular mental state interpretation in each test item.

Second, the notion of mindreading flexibility was explicitly operationalised as a participant's tendency to adjust their ranking of perceived possibility of a mental state interpretation when context was varied in a way that agreed with the majority of fellow test takers. Although flexibility has been suggested to be an essential ability involved in mindreading, it has been rarely studied in the existing literature (e.g., Hayward et al., 2018). Existing work has not examined how varying context influences interpretations as an indicator of flexibility.

Third, both the generation and selection processes were studied without conflation between them, and not solely focusing on the final selection. While existing models such as the ToMM-SP (Leslie et al., 2004) and BToM models (Baker et al., 2017) posit that mindreading involves generation and selection processes, few empirical studies have captured the ways that the generation mechanism differs from the selection process. The current thesis has delineated the two processes in the study of role of context in influencing generation and selection of possible interpretations (Chapter 6).

7.2.2 Overarching implications

This thesis has four overarching implications for theory and future research.

First, there is a need to evaluate the design and administration of existing mindreading measures for neurotypical adults. The findings suggest researchers should rethink the assumption that mindreading can only be measured using a single correct answer, which does not vary by test takers' characteristics or task administration. This is especially relevant to the finding in Chapter 2 that most tasks measure accuracy and some tasks were inconsistently administered across studies. The empirical findings challenge the notion that there is a single correct answer by showing that consensus varies across different groups and task formats. Alignment is suggested as an alternative to accuracy, which is consistent with the view that mindreading is inherently social and that success can be characterised by group agreement

(Apperly et al., 2024). However, alignment and flexibility in responding to open-ended ambiguous social stimuli were not found to be reliable indicators in the current studies.

Second, the operationalisation and calculation of alignment scores offer a feasible way to quantify differences in mental state interpretations across different groups. This method can be used to study the double empathy problem, which concerns differences in mindreading between neurotypes (Edey et al., 2016). The alignment score calculation can also be applied to study mindreading differences across various groups, such as cultural differences, to further evaluate the notion of a single "accurate" mental state interpretation that is invariant between groups of individuals.

Third, the current findings support the notion that propensity and accuracy (with alignment studied as an alternative to accuracy in this thesis) should not be conflated. Propensity was found to be uncorrelated with alignment. The respective processes, generation (related to propensity) and selection (related to accuracy) were also found to be influenced by context to different extents. These findings are in line with the idea that these processes involve distinct mechanisms and that propensity and accuracy are independent facets of mindreading (e.g., Carpenter et al., 2016). The findings are also in line with the view that generation and selection are distinct processes (Apperly et al., 2024; Baker et al., 2017; Leslie et al., 2004).

Fourth, the validity of a decontextualised mental state decoding task might be questioned. Chapters 5 and 6 show that context significantly influenced the selection of the most appropriate mental state interpretations. Therefore, simple decoding of mental states from observable cues in a decontextualised manner is likely insufficient for explaining mindreading success. Instead, the ability to interpret social stimuli in a contextually sensitive manner is also important. These results also corroborate findings showing emotion perception from facial expressions is influenced by context (Barrett et al., 2011). The contexts manipulated in the current studies were limited to background information about social interactions and relationships, but other contexts could also involve characteristics of the target, such as personality, neurotype, and culture. As research has shown these factors influence mindreading accuracy and emotion expression (Adams et al., 2010; Brewer et al., 2016; Perez-Zapata et al., 2016), the basis of assessing mindreading with such simple, decontextualised decoding tasks requires reconsideration.

7.3 General limitations and future research directions

7.3.1 General limitations

The first limitation concerns the generalisability of the findings. The empirical studies in this thesis focused on mental state decoding tasks using a limited set of pictorial social stimuli. This limits the applicability of the results to mental state reasoning tasks that require more complex processing, such as understanding the plot and nuanced interactions between characters to infer mental states, and especially ones that solely present verbal vignettes or use more naturalistic stimuli such as movies. Given that mindreading is a multi-faceted construct, the conclusions may not extend to all types of mindreading. For instance, context may be less relevant when assessing children's sequence of acquisition of mental state concepts compared to evaluating older children's or adults' abilities to make mental state inferences in diverse contexts.

Another limitation is that with a focus on mental state interpretations about a target's "thinking" and "feeling", this thesis has not addressed the multidimensionality of mindreading or identify the core abilities that constitute mindreading, despite finding inconsistent interrelations among existing measures in Chapter 2. Some literature suggests that measures claiming to assess mindreading, such as the RMET, might actually measure emotion recognition rather than mindreading (Oakley et al., 2016). Emotion recognition and mindreading can be dissociated on one hand, as mindreading does not necessitate recognising emotions from observations of a target’s expressions, but on the other hand, emotion recognition could be a component of mindreading depending on the definition adopted. Without a clear taxonomy of mindreading abilities, it is challenging to determine if a task genuinely captures mindreading. Previous reviews (e.g., Happé et al., 2017) have proposed schematic illustrations of how socio-cognitive abilities may interrelate, similar to parallel research in executive functions research that has clearly mapped out the shared variance across subdomains of the target construct (e.g., Friedman & Miyake, 2017; Miyake et al., 2000).

The use of an open-ended format in Studies 1 to 3 and a bottom-up developed coding scheme also presents a limitation. Despite refining the coding scheme in Study 3 and ensuring inter-rater reliability, participants in a different population (e.g., different in cultural background) might provide interpretations that do not fit the existing mental state categories in the current coding scheme. Additionally, qualitative responses could be ambiguous, making it challenging to assign corresponding mental state categories to particular verbal responses. However, the alignment score calculation, which considers all possible categories involved in a response, reduces the risk of underestimating alignment compared to a majority-wins approach: even if a response does not hit the majority response given the judgment of the coder, the response still has a score higher than zero, if it has been coded on at least one more category.

An additional limitation is that neurotypical adults recruited for the empirical studies were practically defined as non-autistic adults, while individuals with other forms of neurodivergence (e.g., ADHD) may have been included. This inclusion could introduce some variability in the data, but it is unlikely to have caused systematic errors that would undermine the main arguments in the chapters. This is because ASD is the clinical condition with the strongest social relevance, and autistic individuals were screened out from the samples.

7.3.2 Future research directions

An obvious direction for future research is to examine whether similar differences in interpretations between groups, as observed in tasks involving a significant mental state decoding component, also occur in purely mental state reasoning tasks using verbal vignettes, and whether consensus varies between different groups of individuals. Preliminary evidence already suggests that mindreading accuracy is higher when the target is culturally similar, as shown in cross-cultural research using an adapted Strange Stories Task (Perez-Zapata et al., 2016). Future studies could test different groups within the same culture and examine variations in participants' flexibility when the same task is adapted to different cultures or scored based on consensus in different groups (if variation in consensus is found between groups). The alignment paradigm, while not requiring the use of the same stimuli as the current studies, can be adapted to study the double empathy problem (Edey et al., 2016). For example, a group of autistic individuals and a group of neurotypical individuals could be recruited to interpret a standardised set of social stimuli, and the same-group alignment scores could be compared to the crossed-group alignment scores. One group showing higher same-group than crossed-group alignment would suggest differences in interpretations by the two groups.

Another future research direction is to compare the utility of determining appropriateness based on established ground truth (i.e., by asking the target what they actually think or feel using the Interview task; Long et al., 2022) and alignment. There is yet no empirical research comparing these two definitions of mindreading success, while mindreading is expected to have consequences for social outcomes (e.g., Apperly, 2012; Dunn & Cutting, 1999; Dunn et al., 1991; Dunn, & Brophy, 2005; Hughes & Devine, 2015). Hence, it is important to examine if the tasks adopting various definitions for mindreading success do predict positive social outcomes, such as better social functioning skills, to establish criterion-based validity of the tasks. A simple way to conduct such research is to adopt a task that is based on established ground truth (e.g. the Interview Task; Long et al., 2022), and score it in two ways: first, according to the known ground truth and second, based on alignment with others. Regression models can be specified to model how well these scores predict social outcomes.

A third future research direction is to map out the taxonomy of mindreading abilities and test competing models using a latent variable approach. A strong theoretical model is lacking in defining the essential nature of mindreading and whether a unitary concept of mindreading truly exists in the adult population, despite attempts in younger populations (e.g., Devine et al., 2023; Hughes et al., 2018). Empirical research in this direction can be conducted by recruiting participants to complete a battery of mindreading tasks and specify different latent variable models to compare the model fit. To validate sub-components of mindreading, tasks should load on distinct latent factors for each sub-component, demonstrating clear separation from other sub-components, while still being interrelated with other domains, hence providing evidence for a unified construct of mindreading. This approach is similar to research in executive functions, where different subdomains including inhibition, shifting, and updating show both distinctiveness and shared variance (Friedman & Miyake, 2017; Miyake et al., 2000; Rodríguez-Nieto et al., 2022).

A final suggested future research direction is to explore how context affects the generation and selection of mental state interpretations. Researchers could manipulate thinking and response time allowed, and vary the similarity between different contexts. If it takes extra cognitive effort to consider context in the process of generating plausible mental state interpretations, it is possible that participants tend to generate interpretations that are applicable not only to the current context but also to alternative contexts when less time is allowed for generation; the difference in likelihood ratings for the same interpretations between the two contexts would then be expected to be reduced. The difference in likelihood ratings could be modulated by similarity between the two contexts as well. This research direction would help to reveal whether taking context into consideration is cognitively effortful, and whether the generation of mental state interpretations is not constrained by the given context to the extent that these generated interpretations apply even to contexts that are highly dissimilar and irrelevant. These manipulations should provide clearer insights into when and how context impacts mindreading.

7.4 Conclusion

This thesis addressed gaps in adult mindreading research by evaluating measurement challenges, proposing alignment as an alternative to accuracy, examining the distinct processes of generation and selection of mental state interpretations, and exploring possible indicators of individual differences in mindreading in neurotypical adults. Chapter 2 revealed inconsistencies in existing measures and called for more research on evaluating existing measures and developing better measures for mindreading. Chapter 3 to 5 challenged the notion of a singular correct interpretation, showing that the most-agreed or best description perceived varied with task format and context, and individuals could flexibly adjust interpretations with varying context. Chapters 3 and 6 demonstrated that generation and selection were distinct processes, with Chapter 6 showing that context strongly constrained the selection process but less so the generation of interpretations. Finally, the number of mental state interpretations generated was found to be a promising indicator of individual differences in propensity of mindreading.

References

Abell, F., Happé, F., & Frith, U. (2000). Do triangles play tricks? Attribution of mental states to animated shapes in normal and abnormal development. *Cognitive Development, 15*(1), 1–16. https://doi.org/10.1016/S0885-2014(00)00014-9

Abu-Akel, A. M., Wood, S. J., Hansen, P. C., & Apperly, I. A. (2015). Perspective-taking abilities in the balance between autism tendencies and psychosis proneness. *Proceedings of the Royal Society B: Biological Sciences*, *282*(1808). https://doi.org/10.1098/rspb.2015.0563

Achim, A. M., Ouellet, R., Roy, M.-A., & Jackson, P. L. (2012). Mentalizing in first-episode psychosis. *Psychiatry Research*, *196*(2–3), 207–213. https://doi.org/10.1016/j.psychres.2011.10.011

Adams, R. B., Rule, N. O., Franklin, R. G., Wang, E., Stevenson, M. T., Yoshikawa, S., Nomura, M., Sato, W., Kveraga, K., & Ambady, N. (2010). Cross-cultural reading the mind in the eyes: An fMRI investigation. *Journal of Cognitive Neuroscience*, *22*(1), 97–108. https://doi.org/10.1162/jocn.2009.21187

Adolphs, R., Baron-Cohen, S., & Tranel, D. (2002). Impaired Recognition of Social Emotions following Amygdala Damage. *Journal of Cognitive Neuroscience*, *14*(8), 1264–1274. https://doi.org/10.1162/089892902760807258

Alkhaldi, R. S., Sheppard, E., & Mitchell, P. (2019). Is There a Link Between Autistic People Being Perceived Unfavorably and Having a Mind That Is Difficult to Read? *Journal of Autism and Developmental Disorders*, *49*(10), 3973–3982. https://doi.org/10.1007/s10803-019-04101-1

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *The Standards for Educational and Psychological Testing*. American Educational Research Association.

Anastasi, A. (1948). The nature of psychological “traits.” *Psychological Review, 55*(3), 127–138. https://doi.org/10.1037/h0063619

Apperly, I. A. (2010). *Mindreaders: the cognitive basis of “theory of mind.”* Psychology Press.

Apperly, I. A. (2012). What is “theory of mind”? Concepts, cognitive processes and individual differences. *Quarterly Journal of Experimental Psychology*, *65*(5), 825–839. https://doi.org/10.1080/17470218.2012.676055

Apperly, I. A. (2021). Cognitive basis of mindreading in middle childhood and adolescence. In *Theory of Mind in Middle Childhood and Adolescence: Integrating Multiple Perspectives* (pp. 37–54). https://doi.org/10.4324/9780429326899-4

Apperly, I. A., Back, E., Samson, D., & France, L. (2008). The cost of thinking about false beliefs: Evidence from adults’ performance on a non-inferential theory of mind task. *Cognition*, *106*(3), 1093–1108. https://doi.org/10.1016/j.cognition.2007.05.005

Apperly, I. A., Samson, D., Chiavarino, C., & Humphreys, G. W. (2004). Frontal and Temporo-Parietal Lobe Contributions to Theory of Mind: Neuropsychological Evidence from a False-Belief Task with Reduced Language and Executive Demands. *Journal of Cognitive Neuroscience*, *16*(10), 1773–1784. https://doi.org/10.1162/0898929042947928

Apperly, I. A., Samson, D., & Humphreys, G. W. (2009b). Studies of adults can inform accounts of theory of mind development. *Developmental Psychology*, *45*(1), 190–201. https://doi.org/10.1037/a0014098

Apperly, I. A., & Wang, J. J. (2021). Mindreading in adults: Cognitive basis, motivation, and individual differences. In H. J. Ferguson & E. E. F. Bradford (Eds.), *The Cognitive Basis of Social Interaction Across the Lifespan* (pp. 96–116). Oxford University Press. https://doi.org/10.1093/oso/9780198843290.003.0005

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., Devine, R. T., & Butterfill, S. (2024). Mindreading is an Asynchronous Joint Activity: The MAJA Account of Theory of Mind performance, and individual differences. *PsyArXiv*. https://doi.org/10.31234/osf.io/6p95c

Atherton, G., & Cross, L. (2022). Reading the mind in cartoon eyes: Comparing human versus cartoon emotion recognition in those with high and low levels of autistic traits. *Psychological Reports, 125*(3), 1380–1396. https://doi.org/10.1177/0033294120988135

Aviezer, H., Trope, Y., & Todorov, A. (2012). Holistic person processing: Faces with bodies tell the whole story. *Journal of Personality and Social Psychology, 103*(1), 20–37. https://doi.org/10.1037/a0027411

Aykan, S., & Nalçacı, E. (2018). Assessing Theory of Mind by Humor: The Humor Comprehension and Appreciation Test (ToM-HCAT). *Frontiers in Psychology, 9*. https://doi.org/10.3389/fpsyg.2018.01470

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

Baker, C., Saxe, R., & Tenenbaum, J. (2011). Bayesian theory of mind: Modeling joint belief-desire attribution. *Proceedings of the Annual Meeting of the Cognitive Science Society, 33*(33), 2469–2474.

Baksh, R. A., Abrahams, S., Auyeung, B., & MacPherson, S. E. (2018). The Edinburgh Social Cognition Test (ESCoT): Examining the effects of age on a new measure of theory of mind and social norm understanding. *PLoS ONE, 13*(4), 1–16. https://doi.org/10.1371/journal.pone.0195818

Ballespí, S., Vives, J., Sharp, C., Tobar, A., & Barrantes-Vidal, N. (2019). Hypermentalizing in Social Anxiety: Evidence for a Context-Dependent Relationship. *Frontiers in Psychology, 10*. https://doi.org/10.3389/fpsyg.2019.01501

Banerjee, R., Watling, D., & Caputi, M. (2011). Peer Relations and the Understanding of Faux Pas: Longitudinal Evidence for Bidirectional Associations. *Child Development, 82*(6), 1887–1905. https://doi.org/10.1111/j.1467-8624.2011.01669.x

Baron-Cohen, S. (1995). *Mindblindness: An Essay on Autism and Theory of Mind.* The MIT Press. https://doi.org/10.7551/mitpress/4635.001.0001

Baron-Cohen, S., Jolliffe, T., Mortimore, C., & Robertson, M. (1997). Another Advanced Test of Theory of Mind: Evidence from Very High Functioning Adults with Autism or Asperger Syndrome. *Journal of Child Psychology and Psychiatry, 38*(7), 813–822. https://doi.org/10.1111/j.1469-7610.1997.tb01599.x

Baron-Cohen, S., Leslie, A. M., & Frith, U. (1985). Does the autistic child have a "theory of mind" ? *Cognition, 21*(1), 37–46. 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. https://doi.org/10.1023/a:1023035012436

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, 42*(2), 241–251. https://doi.org/10.1111/1469-7610.00715

Barrett, L. F., Mesquita, B., & Gendron, M. (2011). Context in emotion perception. *Current Directions in Psychological Science, 20*(5), 286–290. https://doi.org/10.1177/0963721411422522

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using {lme4}. *Journal of Statistical Software, 67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Beaudoin, C., Leblanc, É., Gagner, C., & Beauchamp, M. H. (2020). Systematic Review and Inventory of Theory of Mind Measures for Young Children. *Frontiers in Psychology, 10*(January). https://doi.org/10.3389/fpsyg.2019.02905

Betz, N., Hoemann, K., & Barrett, L. F. (2019). Words are a context for mental inference. *Emotion, 19*(8), 1463–1477. https://doi.org/10.1037/emo0000510

Bialystok, E., & Senman, L. (2004). Executive Processes in Appearance-Reality Tasks: The Role of Inhibition of Attention and Symbolic Representation. *Child Development, 75*(2), 562–579. https://doi.org/https://dx.doi.org/10.1111/j.1467-8624.2004.00693.x

Bialystok, E., & Viswanathan, M. (2009). Components of executive control with advantages for bilingual children in two cultures. *Cognition, 112*(3), 494–500. https://doi.org/10.1016/j.cognition.2009.06.014

Blair, R. J. R. (2000). Impaired social response reversal: A case of `acquired sociopathy'. *Brain, 123*(6), 1122–1141. https://doi.org/10.1093/brain/123.6.1122

Bora, E., Eryavuz, A., Kayahan, B., Sungu, G., & Veznedaroglu, B. (2006). Social functioning, theory of mind and neurocognition in outpatients with schizophrenia; mental state decoding may be a better predictor of social functioning than mental state reasoning. *Psychiatry Research, 145*(2–3), 95–103. https://doi.org/10.1016/j.psychres.2005.11.003

Bosco, F. M., Colle, L., Fazio, S. De, Bono, A., Ruberti, S., & Tirassa, M. (2009). Th.o.m.a.s.: An exploratory assessment of Theory of Mind in schizophrenic subjects. *Consciousness and Cognition, 18*(1), 306–319. https://doi.org/10.1016/j.concog.2008.06.006

Bradford, E. E. F., Jentzsch, I., & Gomez, J.-C. (2015). From self to social cognition: Theory of Mind mechanisms and their relation to Executive Functioning. *Cognition, 138,* 21–34. https://doi.org/10.1016/j.cognition.2015.02.001

Breen, E. K. (1993). Recall and Recognition Memory in Parkinson’s Disease. *Cortex, 29*(1), 91–102. https://doi.org/10.1016/S0010-9452(13)80214-6

Brewer, N., Young, R. L., & Barnett, E. (2017). Measuring Theory of Mind in Adults with Autism Spectrum Disorder. *Journal of Autism and Developmental Disorders, 47*(7), 1927–1941. https://doi.org/10.1007/s10803-017-3080-x

Brewer, N., Young, R. L., Norris, J. E., Maras, K., Michael, Z., & Barnett, E. (2022). A Quick Measure of Theory of Mind in Autistic Adults: Decision Accuracy, Latency and Self-Awareness. *Journal of Autism and Developmental Disorders, 52*(6), 2479–2496. https://doi.org/10.1007/s10803-021-05166-7

Brewer, R., Biotti, F., Catmur, C., Press, C., Happé, 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

Brüne, M. (2003). Theory of mind and the role of IQ in chronic disorganized schizophrenia. *Schizophrenia Research, 60*(1), 57–64. https://doi.org/10.1016/S0920-9964(02)00162-7

Brüne, M. (2005). Emotion recognition, 'theory of mind,' and social behavior in schizophrenia. *Psychiatry Research, 133*(2–3), 135–147. https://doi.org/10.1016/j.psychres.2004.10.007

Brunet, E., Sarfati, Y., Hardy-Baylé, M.-C., & Decety, J. (2000). A PET Investigation of the Attribution of Intentions with a Nonverbal Task. *NeuroImage, 11*(2), 157–166. https://doi.org/10.1006/nimg.1999.0525

Bürkner, P.-C. (2017). brms : An R Package for Bayesian Multilevel Models Using Stan. *Journal of Statistical Software, 80*(1). https://doi.org/10.18637/jss.v080.i01

Bürkner, P.-C. (2018). Advanced Bayesian Multilevel Modeling with the R Package brms. *The R Journal, 10*(1), 395. https://doi.org/10.32614/RJ-2018-017

Calev, A. (1984). Recall and recognition in chronic nondemented schizophrenics: Use of matched tasks. *Journal of Abnormal Psychology, 93*(2), 172–177. https://doi.org/10.1037/0021-843X.93.2.172

Calso, C., Besnard, J., & Allain, P. (2019). Frontal Lobe Functions in Normal Aging: Metacognition, Autonomy, and Quality of Life. *Experimental Aging Research, 45*(1), 10–27. https://doi.org/10.1080/0361073X.2018.1560105

Cantor, N., Mischel, W., & Schwartz, J. C. (1982). A prototype analysis. *Cognitive Psychology, 14*, 45–77.

Canty, A. L., Neumann, D. L., Fleming, J., & Shum, D. H. K. (2017). Evaluation of a newly developed measure of theory of mind: The virtual assessment of mentalising ability. *Neuropsychological Rehabilitation, 27*(5), 834–870. https://doi.org/10.1080/09602011.2015.1052820

Carpenter, J. M., Green, M. C., & Vacharkulksemsuk, T. (2016). Beyond perspective-taking: Mind-reading motivation. *Motivation and Emotion, 40*(3), 358–374. https://doi.org/10.1007/s11031-016-9544-z

Carroll, J. M., & Russell, J. A. (1996). Do facial expressions signal specific emotions? Judging emotion from the face in context. *Journal of Personality and Social Psychology*, *70*(2), 205–218. https://doi.org/10.1037//0022-3514.70.2.205

Cassels, T. G., & Birch, S. A. J. (2014). Comparisons of an Open-Ended vs. Forced-Choice 'Mind Reading' Task: Implications for Measuring Perspective-Taking and Emotion Recognition. *PLoS ONE, 9*(12), e93653. https://doi.org/10.1371/journal.pone.0093653

Castelli, F., Happé, F., Frith, U., & Frith, C. (2000). Movement and Mind: A Functional Imaging Study of Perception and Interpretation of Complex Intentional Movement Patterns. *NeuroImage*, *12*(3), 314–325. https://doi.org/10.1006/nimg.2000.0612

Channon, S., & Crawford, S. (2000). The effects of anterior lesions on performance on a story comprehension test: left anterior impairment on a theory of mind-type task. *Neuropsychologia, 38*(7), 1006–1017. https://doi.org/10.1016/S0028-3932(99)00154-2

Chen, H., Cohen, P., & Chen, S. (2010). How big is a big odds ratio? Interpreting the magnitudes of odds ratios in epidemiological studies. *Communications in Statistics: Simulation and Computation, 39*(4), 860–864. https://doi.org/10.1080/03610911003650383

Cho, I., Kamkar, N., & Hosseini-Kamkar, N. (2022). Reasoning about mental states under uncertainty. *PLOS ONE, 17*(11), e0277356. https://doi.org/10.1371/journal.pone.0277356

Cicchetti, D. V. (1994). Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological Assessment, 6*(4), 284–290. https://doi.org/10.1037/1040-3590.6.4.284

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112*(1), 155–159. https://doi.org/10.1037/0033-2909.112.1.155

Contreras-Huerta, L. S., Pisauro, M. A., & Apps, M. A. J. (2020). Effort shapes social cognition and behaviour: A neuro-cognitive framework. *Neuroscience and Biobehavioral Reviews, 118*(July), 426–439. https://doi.org/10.1016/j.neubiorev.2020.08.003

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

Corcoran, R., Cahill, C., & Frith, C. (1997). The appreciation of visual jokes in people with schizophrenia: a study of 'mentalizing' ability. *Schizophrenia Research, 24*(3), 319–327. https://doi.org/10.1016/S0920-9964(96)00117-X

Corcoran, R., Mercer, G., & Frith, C. D. (1995). Schizophrenia, symptomatology and social inference: Investigating "theory of mind" in people with schizophrenia. *Schizophrenia Research, 17*(1), 5–13. https://doi.org/10.1016/0920-9964(95)00024-G

Cortina, J. M. (1993). What Is Coefficient Alpha? An Examination of Theory and Applications. *Journal of Applied Psychology, 78*(1), 98–104. https://doi.org/10.1037/0021-9010.78.1.98

Dennett, D. C. (1987). *The Intentional Stance*. The MIT Press.

Derksen, D. G., Hunsche, M. C., Giroux, M. E., Connolly, D. A., & Bernstein, D. M. (2018). A systematic review of theory of mind's precursors and functions. Zeitschrift Fur Psychologie / Journal of Psychology, 226(2), 87–97. https://doi.org/10.1027/2151-2604/a000325

Devine, R. T. (2021). Individual differences in theory of mind in middle childhood and adolescence. In *Theory of Mind in Middle Childhood and Adolescence* (pp. 55–76). Routledge. https://doi.org/10.4324/9780429326899-5

Devine, R. T., & Apperly, I. A. (2022). Willing and able? Theory of mind, social motivation, and social competence in middle childhood and early adolescence. *Developmental Science, 25*(1), 1–14. https://doi.org/10.1111/desc.13137

Devine, R. T., & Hughes, C. (2013). Silent Films and Strange Stories: Theory of Mind, Gender, and Social Experiences in Middle Childhood. *Child Development, 84*(3), 989–1003. https://doi.org/10.1111/cdev.12017

Devine, R. T., & Hughes, C. (2016). Measuring theory of mind across middle childhood: Reliability and validity of the Silent Films and Strange Stories tasks. *Journal of Experimental Child Psychology, 149*, 23–40. https://doi.org/10.1016/j.jecp.2015.07.011

Devine, R. T., Kovatchev, V., Traynor, I. G., Smith, P., & Lee, M. (2023). Machine Learning and Deep Learning Systems for Automated Measurement of “Advanced” Theory of Mind: Reliability and Validity in Children and Adolescents. *Psychological Assessment, 35*(2), 165–177. https://doi.org/10.1037/pas0001186

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

Dewey, M. (1991). Living with Asperger’s syndrome. In *Autism and Asperger Syndrome* (pp. 184–206). Cambridge University Press. https://doi.org/10.1017/CBO9780511526770.006

Diaz, V., & Farrar, M. J. (2018). Do bilingual and monolingual preschoolers acquire false belief understanding similarly? The role of executive functioning and language. *First Language, 38*(4), 382–398. https://doi.org/10.1177/0142723717752741

Dodell-Feder, D., Lincoln, S. H., Coulson, J. P., & Hooker, C. I. (2013). Using fiction to assess mental state understanding: A new task for assessing theory of mind in adults. *PLoS ONE*, *8*(11), 1–14. https://doi.org/10.1371/journal.pone.0081279

Dunn, J., & Brophy, M. (2005). Communication, Relationships, and Individual Differences in Children's Understanding of Mind. In *Why Language Matters for Theory of Mind* (pp. 50–69).

Oxford University Press. https://doi.org/10.1093/acprof:oso/9780195159912.003.0003

Dunn, J., Brown, J., Slomkowski, C., Tesla, C., & Youngblade, L. (1991). Young Children ' s Understanding of Other People ' s Feelings and Beliefs : Individual Differences and Their Antecedents. *Child Development, 62*(6), 1352–1366.

Dunn, J., & Cutting, A. L. (1999). Understanding Others, and Individual Differences in Friendship Interactions in Young Children. *Social Development, 8*(2), 201–219. https://doi.org/10.1111/1467-9507.00091

Dwyer, K., David, A. S., McCarthy, R., McKenna, P., & Peters, E. (2020). Linguistic alignment and theory of mind impairments in schizophrenia patients' dialogic interactions. *Psychological Medicine, 50*(13), 2194–2202. https://doi.org/10.1017/S0033291719002289

Dyck, M. J., Ferguson, K., & Shochet, I. M. (2001). Do autism spectrum disorders differ from each other and from non-spectrum disorders on emotion recognition tests? *European Child & Adolescent Psychiatry, 10*(2), 105–116. https://doi.org/10.1007/s007870170033

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

Dziobek, I., Rogers, K., Fleck, S., Bahnemann, M., Heekeren, H. R., Wolf, O. T., & Convit, A. (2008). Dissociation of Cognitive and Emotional Empathy in Adults with Asperger Syndrome Using the Multifaceted Empathy Test (MET). *Journal of Autism and Developmental Disorders, 38*(3), 464–473. https://doi.org/10.1007/s10803-007-0486-x

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

Eickers, G. (2024). Scripts and Social Cognition. *Ergo: an Open Access Journal of Philosophy, 10*(0), 1565–1587. https://doi.org/10.3998/ergo.5191

El Haj, M., Antoine, P., & Nandrino, J. L. (2017). When deception influences memory: the implication of theory of mind. *Quarterly Journal of Experimental Psychology, 70*(7), 1166–1173. https://doi.org/10.1080/17470218.2016.1173079

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. https://doi.org/10.3758/BRM.41.4.1149

Festinger, L. (1957). *A theory of cognitive dissonance.* Stanford University Press.

Fleiss, J. L. (1986). *The design and analysis of clinical experiments*. Wiley.

Fodor, J. A. (1990). *A theory of content and other essays*. The MIT Press.

Friedman, N. P., & Miyake, A. (2017). Unity and diversity of executive functions: Individual differences as a window on cognitive structure. *Cortex, 86,* 186–204. https://doi.org/10.1016/j.cortex.2016.04.023

Frith, C. D., & Corcoran, R. (1996). Exploring 'theory of mind' in people with schizophrenia. *Psychological Medicine*, *26*(3), 521–530. https://doi.org/10.1017/S0033291700035601

Fu, I., Chen, K., Liu, M., Jiang, D., Hsieh, C.-L., & Lee, S.-C. (2023). A systematic review of measures of theory of mind for children. *Developmental Review, 67*(1), 101061. https://doi.org/10.1016/j.dr.2022.101061

Gallagher, H. ., Happé, F., Brunswick, N., Fletcher, P. ., Frith, U., & Frith, C. . (2000). Reading the mind in cartoons and stories: an fMRI study of 'theory of mind' in verbal and nonverbal tasks. *Neuropsychologia, 38*(1), 11–21. https://doi.org/10.1016/S0028-3932(99)00053-6

Gallant, C., & Good, D. (2020). Examining the "reading the mind in the eyes test" as an assessment of subtle differences in affective theory of mind after concussion. *The Clinical Neuropsychologist, 34*(2), 296–317. https://doi.org/10.1080/13854046.2019.1612946

German, T. P., & Hehman, J. A. (2006). Representational and executive selection resources in "theory of mind": Evidence from compromised belief-desire reasoning in old age. *Cognition, 101*(1), 129–152. https://doi.org/10.1016/j.cognition.2005.05.007

Gignac, G. E., & Szodorai, E. T. (2016). Effect size guidelines for individual differences researchers. *Personality and Individual Differences, 102*, 74–78. https://doi.org/10.1016/j.paid.2016.06.069

Gilbert, D. T. (1998). Ordinary Psychology. In D. T. Gilbert, S. T. Fiske, & G. Lindzey (Eds.), *The handbook of social psychology* (4th ed., pp. 8–150). McGraw-Hill.

Gilpin, A. R. (1993). Table for conversion of Kendall's Tau to Spearman's Rho within the context of measures of magnitude of effect for meta-analysis. Educational and *Psychological Measurement, 53*(1), 87–92. https://doi.org/10.1177/0013164493053001007

Girardi, A., MacPherson, S. E., & Abrahams, S. (2011). Deficits in emotional and social cognition in amyotrophic lateral sclerosis. *Neuropsychology, 25*(1), 53–65. https://doi.org/10.1037/a0020357

Goetz, P. J. (2003). The effects of bilingualism on theory of mind development. *Bilingualism: Language and Cognition*, *6*(1), S1366728903001007. https://doi.org/10.1017/S1366728903001007

Golan, O., Baron-Cohen, S., & Hill, J. (2006). The Cambridge Mindreading (CAM) Face-Voice Battery: Testing Complex Emotion Recognition in Adults with and without Asperger Syndrome. *Journal of Autism and Developmental Disorders, 36*(2), 169–183. https://doi.org/10.1007/s10803-005-0057-y

Golan, O., Baron-Cohen, S., Hill, J. J., & Golan, Y. (2006). The “Reading the Mind in Films” Task: Complex emotion recognition in adults with and without autism spectrum conditions. *Social Neuroscience, 1*(2), 111–123. https://doi.org/10.1080/17470910600980986

Golan, O., Baron-Cohen, S., Hill, J. J., & Rutherford, M. D. (2007). The ‘Reading the Mind in the Voice’ Test-Revised: A Study of Complex Emotion Recognition in Adults with and Without Autism Spectrum Conditions. *Journal of Autism and Developmental Disorders, 37*(6), 1096–1106. https://doi.org/10.1007/s10803-006-0252-5

Goldman, A. I. (2006). *Simulating minds: The philosophy, psychology, and neuroscience of mindreading*. Oxford University Press.

Gönültaş, S., Selçuk, B., Slaughter, V., Hunter, J. A., & Ruffman, T. (2020). The Capricious Nature of Theory of Mind: Does Mental State Understanding Depend on the Characteristics of the Target? *Child Development, 91*(2), e280–e298. https://doi.org/10.1111/cdev.13223

Gopnik, A., & Astington, J. W. (1988). Children's understanding of representational change and its relation to the understanding of false belief and the appearance-reality distinction. *Child Development, 59*(1), 26–37. https://doi.org/10.1111/j.1467-8624.1988.tb03192.x

Gopnik, A., & Wellman, H. M. (1992). Why the Child's Theory of Mind Really Is a Theory. *Mind & Language, 7*(1–2), 145–171. https://doi.org/10.1111/j.1468-0017.1992.tb00202.x

Gordon, H. L., Baird, A. A., & End, A. (2004). Functional differences among those high and low on a trait measure of psychopathy. *Biological Psychiatry, 56*(7), 516–521. https://doi.org/10.1016/j.biopsych.2004.06.030

Green, P., & Macleod, C. J. (2016). SIMR: An R package for power analysis of generalized linear mixed models by simulation. *Methods in Ecology and Evolution, 7*(4), 493–498. https://doi.org/10.1111/2041-210X.12504

Gregory, C., Lough, S., Stone, V., Erzinclioglu, S., Martin, L., Baron-Cohen, S., & Hodges, J. R. (2002). Theory of mind in patients with frontal variant frontotemporal dementia and Alzheimer's disease: theoretical and practical implications. *Brain, 125*(4), 752–764. https://doi.org/10.1093/brain/awf079

Happé, F., Brownell, H., & Winner, E. (1999). Acquired `theory of mind' impairments following stroke. *Cognition, 70*(3), 211–240. https://doi.org/10.1016/S0010-0277(99)00005-0

Happé, F., Cook, J. L., & Bird, G. (2017). The Structure of Social Cognition: In(ter)dependence of Sociocognitive Processes. *Annual Review of Psychology, 68*(1), 243–267. https://doi.org/10.1146/annurev-psych-010416-044046

Happe, F., & Frith, U. (1996). The neuropsychology of autism. *Brain, 119*(4), 1377–1400. https://doi.org/10.1093/brain/119.4.1377

Happé, F. G. E., Winner, E., & Brownell, H. (1998). The getting of wisdom: Theory of mind in old age. *Developmental Psychology, 34*(2), 358–362. https://doi.org/10.1037/0012-1649.34.2.358

Happe, F. (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). https://doi.org/10.1007/BF02172093

Harkness, K. L., Sabbagh, M. A., Jacobson, J. A., Chowdrey, N. K., & Chen, T. (2005). Enhanced accuracy of mental state decoding in dysphoric college students. *Cognition and Emotion, 19*(7), 999–1025. https://doi.org/10.1080/02699930541000110

Harris, P. L. (1992). From Simulation to Folk Psychology: The Case for Development. *Mind & Language, 7*(1–2), 120–144. https://doi.org/10.1111/j.1468-0017.1992.tb00201.x

Hasson, U., Ghazanfar, A. A., Galantucci, B., Garrod, S., & Keysers, C. (2012). Brain-to-brain coupling: a mechanism for creating and sharing a social world. *Trends in Cognitive Sciences, 16*(2), 114–121. https://doi.org/10.1016/j.tics.2011.12.007

Hayward, E. O., Homer, B. D., & Sprung, M. (2018). Developmental Trends in Flexibility and Automaticity of Social Cognition. *Child Development, 89*(3), 914–928. https://doi.org/10.1111/cdev.12705

Hedge, C., Powell, G., & Sumner, P. (2018). The reliability paradox: Why robust cognitive tasks do not produce reliable individual differences. *Behavior Research Methods, 50*(3), 1166–1186. https://doi.org/10.3758/s13428-017-0935-1

Henry, A., Tourbah, A., Chaunu, M.-P., Rumbach, L., Montreuil, M., & Bakchine, S. (2011). Social Cognition Impairments in Relapsing-Remitting Multiple Sclerosis. *Journal of the International Neuropsychological Society, 17*(6), 1122–1131. https://doi.org/10.1017/S1355617711001147

Hughes, C. (2016). Theory of mind grows up: Reflections on new research on theory of mind in middle childhood and adolescence. *Journal of Experimental Child Psychology, 149,* 1–5. https://doi.org/10.1016/j.jecp.2016.01.017

Hughes, C., Adlam, A., Happé, F., Jackson, J., Taylor, A., & Caspi, A. (2000). Good test-retest reliability for standard and advanced false-belief tasks across a wide range of abilities. *Journal of Child Psychology and Psychiatry and Allied Disciplines, 41*(4), 483–490. https://doi.org/10.1017/S0021963099005533

Hughes, C., & Devine, R. T. (2015). Individual Differences in Theory of Mind From Preschool to Adolescence: Achievements and Directions. *Child Development Perspectives, 9*(3), 149–153. https://doi.org/10.1111/cdep.12124

Hughes, C., Devine, R. T., & Wang, Z. (2018). Does Parental Mind-Mindedness Account for Cross-Cultural Differences in Preschoolers' Theory of Mind? *Child Development, 89*(4), 1296–1310. https://doi.org/10.1111/cdev.12746

Hyman, I. E., & Loftus, E. F. (1998). Errors in autobiographical memory. *Clinical Psychology Review, 18*(8), 933–947. https://doi.org/10.1016/S0272-7358(98)00041-5

Ickes, W. (1993). Empathic Accuracy. *Journal of Personality, 61*(4), 587–610. https://doi.org/10.1111/j.1467-6494.1993.tb00783.x

Ickes, W., Buysse, A., Pham, H., Rivers, K., Erickson, J. R., Hancock, M., Kelleher, J., & Gesn, P. R. (2000). On the difficulty of distinguishing "good" and "poor" perceivers: A social relations analysis of empathic accuracy data. *Personal Relationships, 7*(2), 219–234. https://doi.org/10.1111/j.1475-6811.2000.tb00013.x

Ickes, W., Robertson, E., Tooke, W., & Teng, G. (1986). Naturalistic social cognition: Methodology, assessment, and validation. *Journal of Personality and Social Psychology, 51*(1), 66–82. https://doi.org/10.1037/0022-3514.51.1.66

Imuta, K., Henry, J. D., Slaughter, V., Selcuk, B., & Ruffman, T. (2016). Theory of mind and prosocial behavior in childhood: A meta-analytic review. *Developmental Psychology, 52*(8), 1192–1205. https://doi.org/10.1037/dev0000140

Jack, R. E., Garrod, O. G. B., Yu, H., Caldara, R., & Schyns, P. G. (2012). Facial expressions of emotion are not culturally universal. *Proceedings of the National Academy of Sciences, 109*(19), 7241–7244. https://doi.org/10.1073/pnas.1200155109

Jacoby, L. L., Craik, F. I. M., & Begg, I. (1979). Effects of decision difficulty on recognition and recall. *Journal of Verbal Learning and Verbal Behavior*, *18*(5), 585–600. https://doi.org/10.1016/S0022-5371(79)90324-4

Kanske, P., Böckler, A., Trautwein, F. M., & Singer, T. (2015). Dissecting the social brain: Introducing the EmpaToM to reveal distinct neural networks and brain-behavior relations for empathy and Theory of Mind. *NeuroImage, 122*, 6–19. https://doi.org/10.1016/j.neuroimage.2015.07.082

Karmakar, A., & Dogra, A. K. (2019). Assessment of Theory of Mind in Adults: Beyond False Belief Tasks. *Activitas Nervosa Superior, 61*(3), 142–146. https://doi.org/10.1007/s41470-019-00028-1

Kéri, S., Kállai, I., & Csigó, K. (2020). Attribution of Mental States in Glossolalia: A Direct Comparison With Schizophrenia. *Frontiers in Psychology, 11*. https://doi.org/10.3389/fpsyg.2020.00638

Keysar, B., Barr, D. J., Balin, J. A., & Brauner, J. S. (2000). Taking perspective in conversation: The role of mutual knowledge in comprehension. *Psychological Science, 11*(1), 32–38. https://doi.org/10.1111/1467-9280.00211

Keysar, B., Lin, S., & Barr, D. J. (2003). Limits on theory of mind use in adults. *Cognition, 89*, 25–41. https://doi.org/10.1016/S0010-0277(03)00064-7

Killen, M., Lynn Mulvey, K., Richardson, C., Jampol, N., & Woodward, A. (2011). The accidental transgressor: Morally-relevant theory of mind. *Cognition, 119*(2), 197–215. https://doi.org/10.1016/j.cognition.2011.01.006

Kimball, A. E., Shantz, K., Eager, C., & Roy, J. (2019). Confronting Quasi-Separation in Logistic Mixed Effects for Linguistic Data: A Bayesian Approach. *Journal of Quantitative Linguistics, 26*(3), 231–255. https://doi.org/10.1080/09296174.2018.1499457

Kinderman, P., Dunbar, R., & Bentall, R. P. (1998). Theory-of-mind deficits and causal attributions. *British Journal of Psychology, 89*(2), 191–204. https://doi.org/10.1111/j.2044-8295.1998.tb02680.x

Klin, A. (2000). Attributing Social Meaning to Ambiguous Visual Stimuli in Higher-functioning Autism and Asperger Syndrome: The Social Attribution Task. *Journal of Child Psychology and Psychiatry, 41*(7), 831–846. https://doi.org/10.1111/1469-7610.00671

Kloo, D., & Perner, J. (2003). Training Transfer Between Card Sorting and False Belief Understanding: Helping Children Apply Conflicting Descriptions. *Child Development, 74*(6), 1823–1839. https://doi.org/10.1046/j.1467-8624.2003.00640.x

Kocsis-Bogár, K., Kotulla, S., Maier, S., Voracek, M., & Hennig-Fast, K. (2017). Cognitive Correlates of Different Mentalizing Abilities in Individuals with High and Low Trait Schizotypy: Findings from an Extreme-Group Design. *Frontiers in Psychology, 8*. https://doi.org/10.3389/fpsyg.2017.00922

Kosmidis, M. H., Giannakou, M., Garyfallos, G., Kiosseoglou, G., & Bozikas, V. P. (2011). The Impact of Impaired "Theory of Mind" on Social Interactions in Schizophrenia. *Journal of the International Neuropsychological Society, 17*(3), 511–521. https://doi.org/10.1017/S1355617711000300

Koster-Hale, J., Dodell-Feder, D., & Saxe, R. (2012). [unpublished instrument].

Kovács, Á. M. (2009). Early bilingualism enhances mechanisms of false-belief reasoning. *Developmental Science, 12*(1), 48–54. https://doi.org/10.1111/j.1467-7687.2008.00742.x

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). {lmerTest} Package: Tests in Linear Mixed Effects Models. *Journal of Statistical Software, 82*(13), 1–26. https://doi.org/10.18637/jss.v082.i13

Lagattuta, K. H., Elrod, N. M., & Kramer, H. J. (2016). How do thoughts, emotions, and decisions align? A new way to examine theory of mind during middle childhood and beyond. *Journal of Experimental Child Psychology, 149*, 116–133. https://doi.org/10.1016/j.jecp.2016.01.013

Lakens, D. (2022). Improving Your Statistical Inferences. Retrieved from https://lakens.github.io/statistical_inferences/. https://doi.org/10.5281/zenodo.6409077

Langdon, R., Coltheart, M., Ward, P. B., & Catts, S. V. (2002). Disturbed communication in schizophrenia: the role of poor pragmatics and poor mind-reading. *Psychological Medicine*, *32*(7), 1273–1284. https://doi.org/10.1017/S0033291702006396

Langdon, Robyn, Michie, P. T., Ward, P. B., McConaghy, N., Catts, S. V., & Coltheart, M. (1997). Defective Self and/or Other Mentalising in Schizophrenia: A Cognitive Neuropsychological Approach. *Cognitive Neuropsychiatry, 2*(3), 167–193. https://doi.org/10.1080/135468097396324

Lawson, J., Baron-Cohen, S., & Wheelwright, S. (2004). Empathising and Systemising in Adults with and without Asperger Syndrome. *Journal of Autism and Developmental Disorders, 34*(3), 301–310. https://doi.org/10.1023/B:JADD.0000029552.42724.1b

Leslie, A. M., Friedman, O., & German, T. P. (2004). Core mechanisms in "theory of mind." *Trends in Cognitive Sciences, 8*(12). https://doi.org/10.1016/j.tics.2004.10.001

Licata, M., Zietlow, A.-L., Träuble, B., Sodian, B., & Reck, C. (2016). Maternal Emotional Availability and Its Association with Maternal Psychopathology, Attachment Style Insecurity and Theory of Mind. *Psychopathology, 49*(5), 334–340. https://doi.org/10.1159/000447781

Liu, F., & Eugenio, E. C. (2018). A review and comparison of Bayesian and likelihood-based inferences in beta regression and zero-or-one-inflated beta regression. *Statistical Methods in Medical Research, 27*(4), 1024–1044. https://doi.org/10.1177/0962280216650699

Livingston, L. A., Shah, P., White, S. J., & Happé, F. (2021). Further developing the <scp>Frith–Happé</scp> animations: A quicker, more objective, and web-based test of theory of mind for autistic and neurotypical adults. *Autism Research, 14*(9), 1905–1912. https://doi.org/10.1002/aur.2575

Lockwood, P. L., Ang, Y.-S., Husain, M., & Crockett, M. J. (2017). Individual differences in empathy are associated with apathy-motivation. *Scientific Reports, 7*(1), 17293. https://doi.org/10.1038/s41598-017-17415-w

Long, E. L., Cuve, H. C., Conway, J. R., Catmur, C., & Bird, G. (2022). Novel theory of mind task demonstrates representation of minds in mental state inference. *Scientific Reports, 12*(1), 1–14. https://doi.org/10.1038/s41598-022-25490-x

Loth, E., Gómez, J. C., & Happé, F. (2008). Event schemas in autism spectrum disorders: The role of theory of mind and weak central coherence. *Journal of Autism and Developmental Disorders, 38*(3), 449–463. https://doi.org/10.1007/s10803-007-0412-2

MacCann, C., & Roberts, R. D. (2008). New paradigms for assessing emotional intelligence: Theory and data. *Emotion, 8*(4), 540–551. https://doi.org/10.1037/a0012746

Mahy, C. E. V., Moses, L. J., & Pfeifer, J. H. (2014). How and where: Theory-of-mind in the brain. *Developmental Cognitive Neuroscience, 9*, 68–81. https://doi.org/10.1016/j.dcn.2014.01.002

Mangiafico, S. (2022). rcompanion: Functions to Support Extension Education Program Evaluation. R package version 2.4.15. https://doi.org/10.32614/CRAN.package.rcompanion

Matheson, G. J. (2019). We need to talk about reliability: making better use of test-retest studies for study design and interpretation. *PeerJ, 7*(5), e6918. https://doi.org/10.7717/peerj.6918

Mayes, L. C., Klin, A., Tercyak, K. P., Cicchetti, D. V, & Cohen, D. J. (1996). Test-Retest Reliability for False-Belief Tasks. *Journal of Child Psychology and Psychiatry, 37*(3), 313–319. https://doi.org/10.1111/j.1469-7610.1996.tb01408.x

McDonald, S., Flanagan, S., Rollins, J., & Kinch, J. (2003). TASIT: A New Clinical Tool for Assessing Social Perception After Traumatic Brain Injury. *Journal of Head Trauma Rehabilitation, 18*(3), 219–238. https://doi.org/10.1097/00001199-200305000-00001

McGarry, K. A., West, M., & Hogan, K. F. (2021). Perspective-taking and social competence in adults. *Advances in Cognitive Psychology, 17*(2), 129–135. https://doi.org/10.5709/ACP-0323-5

McGlade, N., Behan, C., Hayden, J., O'Donoghue, T., Peel, R., Haq, F., Gill, M., Corvin, A., O'Callaghan, E., & Donohoe, G. (2008). Mental state decoding v. mental state reasoning as a mediator between cognitive and social function in psychosis. *British Journal of Psychiatry, 193*(1), 77–78. https://doi.org/10.1192/bjp.bp.107.044198

Mehta, J., Starmer, C., & Sugden, R. (1994). The Nature of Salience: An Experimental Investigation of Pure Coordination Games. *The American Economic Review, 84*(3), 658–673. http://dx.doi.org/10.1016/j.jaci.2012.05.050

Meinhardt-Injac, B., Daum, M. M., & Meinhardt, G. (2020). Theory of mind development from adolescence to adulthood: Testing the two-component model. *British Journal of Developmental Psychology*, bjdp.12320. https://doi.org/10.1111/bjdp.12320

Mendoza, J. L., & Mumford, M. (1987). Corrections for Attenuation and Range Restriction on the Predictor. *Journal of Educational Statistics, 12*(3), 282. https://doi.org/10.2307/1164688

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

Mitchell, R. L. C., & Phillips, L. H. (2015). The overlapping relationship between emotion perception and theory of mind. *Neuropsychologia, 70*, 1–10. https://doi.org/10.1016/j.neuropsychologia.2015.02.018

Miyake, A., Friedman, N. P., Emerson, M. J., Witzki, A. H., Howerter, A., & Wager, T. D. (2000). The Unity and Diversity of Executive Functions and Their Contributions to

Complex “Frontal Lobe” Tasks: A Latent Variable Analysis. *Cognitive Psychology, 41*(1), 49–100. https://doi.org/10.1006/cogp.1999.0734

Mordal, J., Gundersen, Ø., & Bramness, J. G. (2010). Norwegian version of the Mini-International Neuropsychiatric Interview: Feasibility, acceptability and test-retest reliability in an acute psychiatric ward. *European Psychiatry, 25*(3), 172–177. https://doi.org/10.1016/j.eurpsy.2009.02.004

Morris, A., Phillips, J., Huang, K., & Cushman, F. (2021). Generating Options and Choosing Between Them Depend on Distinct Forms of Value Representation. *Psychological Science, 32*(11), 1731–1746. https://doi.org/10.1177/09567976211005702

Murphy, B. A., & Lilienfeld, S. O. (2019). Are self-report cognitive empathy ratings valid proxies for cognitive empathy ability? Negligible meta-analytic relations with behavioral task performance. *Psychological Assessment, 31*(8), 1062–1072. https://doi.org/10.1037/pas0000732

Murray, K., Johnston, K., Cunnane, H., Kerr, C., Spain, D., Gillan, N., Hammond, N., Murphy, D., & 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

Navarro, E., Goring, S. A., & Conway, A. R. A. (2021). The relationship between theory of mind and intelligence: a formative g approach. *Journal of Intelligence, 9*(1), 1–15. https://doi.org/10.3390/jintelligence9010011

Nilsen, E. S., & Duong, D. (2013). Depressive symptoms and use of perspective taking within a communicative context. *Cognition and Emotion, 27*(2), 335–344. https://doi.org/10.1080/02699931.2012.708648

Nunally, J. C. (1978). Psychometric Theory (2nd ed.). McGraw-Hill.

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

Orbelo, D. M., Grim, M. A., Talbott, R. E., & Ross, E. D. (2005). Impaired Comprehension of Affective Prosody in Elderly Subjects Is Not Predicted by Age-Related Hearing Loss or Age-Related Cognitive Decline. *Journal of Geriatric Psychiatry and Neurology, 18*(1), 25–32. https://doi.org/10.1177/0891988704272214

Osterhaus, C., & Bosacki, S. L. (2022). Looking for the lighthouse: A systematic review of advanced theory-of-mind tests beyond preschool. *Developmental Review, 64*(February), 101021. https://doi.org/10.1016/j.dr.2022.101021

Ouellet, J., Scherzer, P. B., Rouleau, I., Métras, P., Bertrand-Gauvin, C., Djerroud, N., Boisseau, É., & Duquette, P. (2010). Assessment of social cognition in patients with multiple sclerosis. *Journal of the International Neuropsychological Society, 16*(2), 287–296. https://doi.org/10.1017/S1355617709991329

Paal, T., & Bereczkei, T. (2007). Adult theory of mind, cooperation, Machiavellianism: The effect of mindreading on social relations. *Personality and Individual Differences, 43*(3), 541–551. https://doi.org/10.1016/j.paid.2006.12.021

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., … Moher, D. (2021). The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *BMJ, 372*, n71. https://doi.org/10.1136/bmj.n71

Paal, T., & Bereczkei, T. (2007). Adult theory of mind, cooperation, Machiavellianism: The effect of mindreading on social relations. *Personality and Individual Differences, 43*(3), 541–551. https://doi.org/10.1016/j.paid.2006.12.021

Pellicano, E., & den Houting, J. (2022). Annual Research Review: Shifting from 'normal science' to neurodiversity in autism science. *Journal of Child Psychology and Psychiatry, 63*(4), 381–396. https://doi.org/10.1111/jcpp.13534

Perez-Zapata, D. I. (2023). *The Cognitive Basis of Alignment of Intuitions* [Doctoral dissertation, University of Birmingham]. UBIRA E THESES. http://etheses.bham.ac.uk/id/eprint/13794

Perez-Zapata, D. I., & Apperly, I. (2022). An International Study of Pure Coordination Games: Adaptable Solutions When Intuitions are Presumed to Vary. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.4295474

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

Perner, J. (1991). *Understanding the representational mind*. MIT Press. https://doi.org/10.1016/s0191-6599(96)90063-7

Perner, J., Leekam, S. R., & Wimmer, H. (1987). Three-year-olds' difficulty with false belief: The case for a conceptual deficit. *British Journal of Developmental Psychology, 5*(2), 125–137. https://doi.org/10.1111/j.2044-835x.1987.tb01048.x

Perner, J., & Wimmer, H. (1985). "John thinks that Mary thinks that…" attribution of second-order beliefs by 5- to 10-year-old children. *Journal of Experimental Child Psychology, 39*(3), 437–471. https://doi.org/10.1016/0022-0965(85)90051-7

Petersen, I. T., Hoyniak, C. P., McQuillan, M. E., Bates, J. E., & Staples, A. D. (2016). Measuring the development of inhibitory control: The challenge of heterotypic continuity. *Developmental Review, 40*(3), 25–71. https://doi.org/10.1016/j.dr.2016.02.001

Peterson, C. C., Wellman, H. M., & Slaughter, V. (2012). The Mind Behind the Message: Advancing Theory-of-Mind Scales for Typically Developing Children, and Those With Deafness, Autism, or Asperger Syndrome. *Child Development, 83*(2), 469–485. https://doi.org/10.1111/j.1467-8624.2011.01728.x

Peyroux, E., Prost, Z., Danset-Alexandre, C., Brenugat-Herne, L., Carteau-Martin, I., Gaudelus, B., Jantac, C., Attali, D., Amado, I., Graux, J., Houy-Durand, E., Plasse, J., & Franck, N. (2019). From "under" to "over" social cognition in schizophrenia: Is there distinct profiles of impairments according to negative and positive symptoms? Schizophrenia Research: *Cognition, 15,* 21–29. https://doi.org/10.1016/j.scog.2018.10.001

Pickup, G. J., & Frith, C. D. (2001). Theory of mind impairments in schizophrenia: symptomatology, severity and specificity. *Psychological Medicine, 31*(2), 207–220. https://doi.org/10.1017/S0033291701003385

Pisani, S., Murphy, J., Conway, J., Millgate, E., Catmur, C., & Bird, G. (2021). The relationship between alexithymia and theory of mind: A systematic review. *Neuroscience and Biobehavioral Reviews, 131,* 497–524. https://doi.org/10.1016/j.neubiorev.2021.09.036

Poletti, M., Enrici, I., & Adenzato, M. (2012). Cognitive and affective Theory of Mind in neurodegenerative diseases: Neuropsychological, neuroanatomical and neurochemical levels. *Neuroscience & Biobehavioral Reviews, 36*(9), 2147–2164. https://doi.org/10.1016/j.neubiorev.2012.07.004

Pomareda, C. (2023). Individual Differences in Adults' Mindreading: Psychometric Challenges and the Role of Social Motivation [Unpublished doctoral dissertation]. University of Birmingham.

Premack, D., & Woodruff, G. (1978). Does the chimpanzee have a theory of mind? *Behavioral and Brain Sciences, 1*(4), 515–526. https://doi.org/10.1017/S0140525X00076512

Quesque, F., & Rossetti, Y. (2020). What Do Theory-of-Mind Tasks Actually Measure? Theory and Practice. *Perspectives on Psychological Science, 15*(2), 384–396. https://doi.org/10.1177/1745691619896607

Qureshi, A. W., Monk, R. L., Samson, D., & Apperly, I. A. (2020). Does interference between self and other perspectives in theory of mind tasks reflect a common underlying process? Evidence from individual differences in theory of mind and inhibitory control. *Psychonomic Bulletin and Review, 27*(1), 178–190. https://doi.org/10.3758/s13423-019-01656-z

R Core Team. (2021). R: A Language and Environment for Statistical Computing. https://www.r-project.org/

Rakoczy, H., Wandt, R., Thomas, S., Nowak, J., & Kunzmann, U. (2018). Theory of mind and wisdom: The development of different forms of perspective-taking in late adulthood. *British Journal of Psychology, 109*(1), 6–24. https://doi.org/10.1111/bjop.12246

Revelle, W., & Condon, D. M. (2019). Reliability from α to ω: A tutorial. *Psychological Assessment, 31*(12), 1395–1411. https://doi.org/10.1037/pas0000754

Rice, K., & Redcay, E. (2015). Spontaneous mentalizing captures variability in the cortical thickness of social brain regions. *Social Cognitive and Affective Neuroscience, 10*(3), 327–334. https://doi.org/10.1093/scan/nsu081

Richard, F. D., Bond, C. F., & Stokes-Zoota, J. J. (2003). One Hundred Years of Social Psychology Quantitatively Described. *Review of General Psychology*, *7*(4), 331–363. https://doi.org/10.1037/1089-2680.7.4.331

Richell, R. ., Mitchell, D. G. ., Newman, C., Leonard, A., Baron-Cohen, S., & Blair, R. J. . (2003). Theory of mind and psychopathy: can psychopathic individuals read the 'language of the eyes'? *Neuropsychologia*, *41*(5), 523–526. https://doi.org/10.1016/S0028-3932(02)00175-6

Roberts, D. L., Fiszdon, J., & Tek, C. (2011). Ecological validity of the social cognition screening questionnaire (SCSQ). *Abstracts for the 13th International Congress on Schizophrenia Research (ICOSR)*, 280. https://doi.org/10.1093/schbul/sbq173

Rodríguez-Nieto, G., Seer, C., Sidlauskaite, J., Vleugels, L., Van Roy, A., Hardwick, R., & Swinnen, S. (2022). Inhibition, Shifting and Updating: Inter and intra-domain commonalities and differences from an executive functions activation likelihood estimation meta-analysis. *NeuroImage, 264*, 119665. https://doi.org/10.1016/j.neuroimage.2022.119665

Rönkkö, M., & Cho, E. (2022). An Updated Guideline for Assessing Discriminant Validity. *Organizational Research Methods, 25*(1), 6–14. https://doi.org/10.1177/1094428120968614

Rosenblau, G., Kliemann, D., Heekeren, H. R., & Dziobek, I. (2015). Approximating Implicit and Explicit Mentalizing with Two Naturalistic Video-Based Tasks in Typical Development and Autism Spectrum Disorder. *Journal of Autism and Developmental Disorders, 45*(4), 953–965. https://doi.org/10.1007/s10803-014-2249-9

Rust, J., Kosinski, M., & Stillwell, D. (2020). *Modern Psychometrics*. Routledge. https://doi.org/10.4324/9781315637686

Rutherford, M. D. (2004). The effect of social role on theory of mind reasoning. *British Journal of Psychology, 95*(1), 91–103. https://doi.org/10.1348/000712604322779488

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

Sarfati, Y., Hardybayle, M., Besche, C., & Widlocher, D. (1997). Attribution of intentions to others in people with schizophrenia: a non-verbal exploration with comic strips. *Schizophrenia Research, 25*(3), 199–209. https://doi.org/10.1016/S0920-9964(97)00025-X

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

Schank, R. C., & Abelson, R. P. (1977). *Scripts, plans, goals, and understanding: An inquiry into human knowledge structures*. Lawrence Erlbaum.

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

Sebastian, C. L., Fontaine, N. M. G., Bird, G., Blakemore, S.-J., De Brito, S. A., McCrory, E. J. P., & Viding, E. (2012). Neural processing associated with cognitive and affective Theory of Mind in adolescents and adults. *Social Cognitive and Affective Neuroscience, 7*(1), 53–63. https://doi.org/10.1093/scan/nsr023

Shah, P., Catmur, C., & Bird, G. (2017). From heart to mind: Linking interoception, emotion, and theory of mind. *Cortex, 93*, 220–223. https://doi.org/10.1016/j.cortex.2017.02.010

Shamay-Tsoory, S. G., Shur, S., Barcai-Goodman, L., Medlovich, S., Harari, H., & Levkovitz, Y. (2007). Dissociation of cognitive from affective components of theory of mind in schizophrenia. *Psychiatry Research, 149*(1–3), 11–23. https://doi.org/10.1016/j.psychres.2005.10.018

Sharp, C. (2008). Theory of Mind and conduct problems in children: Deficits in reading the "emotions of the eyes." *Cognition & Emotion, 22*(6), 1149–1158. https://doi.org/10.1080/02699930701667586

Sharp, C., & Hernandez, J. (2021). Mindreading and psychopathology in middle childhood and adolescence. In R. T. Devine & S. Lecce (Eds.), *Theory of Mind in Middle Childhood and Adolescence* (pp. 231–252). Routledge. https://doi.org/10.4324/9780429326899-15

Shaw, P., Lawrence, E. J., Radbourne, C., Bramham, J., Polkey, C. E., & David, A. S. (2004). The impact of early and late damage to the human amygdala on 'theory of mind' reasoning. *Brain*, *127*(7), 1535–1548. https://doi.org/10.1093/brain/awh168

Sommerville, J. A., Bernstein, D. M., & Meltzoff, A. N. (2013). Measuring Beliefs in Centimeters: Private Knowledge Biases Preschoolers' and Adults' Representation of Others' Beliefs. *Child Development, 84*(6), 1846–1854. https://doi.org/10.1111/cdev.12110

Spaulding, S. (2018). *How We Understand Others: Philosophy and Social Cognition.* Routledge.

Sullivan, S., & Ruffman, T. (2004). Social understanding: How does it fare with advancing years? *British Journal of Psychology*, *95*(1), 1–18. https://doi.org/10.1348/000712604322779424

Tahazadeh, S., Barahmand, U., Yaghooti, F., & Nazari, M. A. (2020). Mind Reading in Films Task to Assess Social Cognitive Deficits in Autism Spectrum Conditions. *Journal of Evidence-Based Psychotherapies, 20*(2), 79–100. https://doi.org/10.24193/jebp.2020.2.13

Tahiroglu, D., & Taylor, M. (2019). Anthropomorphism, social understanding, and imaginary companions. *British Journal of Developmental Psychology, 37*(2), 284–299. https://doi.org/10.1111/bjdp.12272

Taylor, D., Gönül, G., Alexander, C., Züberbühler, K., Clément, F., & Glock, H. (2023). Reading minds or reading scripts? De-intellectualising theory of mind. *Biological Reviews, 12*. https://doi.org/10.1111/brv.12994

Tonks, J., Williams, W. H., Frampton, I., Yates, P., & Slater, A. (2007). Assessing emotion recognition in 9–15-years olds: Preliminary analysis of abilities in reading emotion from faces, voices and eyes. *Brain Injury, 21*(6), 623–629. https://doi.org/10.1080/02699050701426865

Trnka, R., & Smelik, V. (2020). Elimination of bias in introspection: Methodological advances, refinements, and recommendations. *New Ideas in Psychology, 56*(July 2019). https://doi.org/10.1016/j.newideapsych.2019.100753

Tulving, E., & Watkins, M. J. (1973). Continuity between Recall and Recognition. *The American Journal of Psychology, 86*(4), 739. https://doi.org/10.2307/1422081

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

Wang, X., Su, Y., Pei, M., & Hong, M. (2021). How self-other control determines individual differences in adolescents' theory of mind. *Cognitive Development, 57*, 101007. https://doi.org/10.1016/j.cogdev.2021.101007

Wang, Z., Devine, R. T., Wong, K. K., & Hughes, C. (2016). Theory of mind and executive function during middle childhood across cultures. *Journal of Experimental Child Psychology, 149*, 6–22. https://doi.org/10.1016/j.jecp.2015.09.028

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

Watson, A. C., Nixon, C. L., Wilson, A., & Capage, L. (1999). Social interaction skills and theory of mind in young children. *Developmental Psychology*, *35*(2), 386–391. https://doi.org/10.1037/0012-1649.35.2.386

Waytz, A., Gray, K., Epley, N., & Wegner, D. M. (2010). Causes and consequences of mind perception. *Trends in Cognitive Sciences, 14*(8), 383–388. https://doi.org/10.1016/J.TICS.2010.05.006

Weimer, A. A., Warnell, K. R., Ettekal, I., Cartwright, K. B., Guajardo, N. R., & Liew, J. (2021). Correlates and antecedents of theory of mind development during middle childhood and adolescence: An integrated model. *Developmental Review, 59*(December 2020), 100945. https://doi.org/10.1016/j.dr.2020.100945

Weinstein, N. Y., Whitmore, L. B., & Mills, K. L. (2022). Individual Differences in Mentalizing Tendencies. *Collabra: Psychology, 8*(1), 1–22. https://doi.org/10.1525/collabra.37602

Wellman, H. M. (2018). Theory of mind: The state of the art*. *European Journal of Developmental Psychology, 15*(6), 728–755. https://doi.org/10.1080/17405629.2018.1435413

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

Wimmer, H., & Perner, J. (1983). Beliefs about beliefs: Representation and constraining function of wrong beliefs in young children's understanding of deception. *Cognition, 13*(1), 103–128. https://doi.org/10.1016/0010-0277(83)90004-5

Yirmiya, N., Erel, O., Shaked, M., & Solomonica-Levi, D. (1998). Meta-Analyses Comparing Theory of Mind Abilities of Individuals with Autism, Individuals with Mental Retardation, and Normally Developing Individuals. *Psychological Bulletin, 124*(3), 283–307. https://doi.org/10.1037/0033-2909.124.3.283

Young, L., Cushman, F., Hauser, M., & Saxe, R. (2007). The neural basis of the interaction between theory of mind and moral judgment. *Proceedings of the National Academy of Sciences, 104*(20), 8235–8240. https://doi.org/10.1073/pnas.0701408104

Zacks, J. M. (2020). Event Perception and Memory. *Annual Review of Psychology, 71*(1979), 165–191. https://doi.org/10.1146/annurev-psych-010419-051101

Zaki, J., Bolger, N., & Ochsner, K. (2008). It Takes Two. *Psychological Science, 19*(4), 399–404. https://doi.org/10.1111/j.1467-9280.2008.02099.x

Zaki, J., Bolger, N., & Ochsner, K. (2009). Unpacking the informational bases of empathic accuracy. *Emotion, 9*(4), 478–487. https://doi.org/10.1037/a0016551

Ziatabar Ahmadi, S. Z., Jalaie, S., & Ashayeri, H. (2015). Validity and Reliability of Published Comprehensive Theory of Mind Tests for Normal Preschool Children: A Systematic Review. *Iranian Journal of Psychiatry, 10*(4), 214–224.

Appendices

Appendix A. Coding scheme for coding verbatim responses into categories

Category	Description	Example(s)	Valence
compassionate	States or implies that the character feels for or cares about the feeling of another person and/or wants to offer help.	concerned, empathetic, protective, trying to help, comforting, supportive	positive
calm/relaxed	States or implies that the character is feeling relaxed, comfortable or at ease.	at ease, relaxed, comfortable	positive
confident	States that the character is certain of his/her abilities and/or has power/influence over others/matters.	confident, important, powerful	positive
curious/interested	States or implies that the character is pleasantly eager to know something. (If a positive valence is not implied, code as attentive/focused/engaged)	curious, interested, intrigued	positive
happy	States or implies that the character is in a state of pleasure, contentment or feels fortunate.	happy, enjoying, content, lucky, pleased, amused	positive
romantic feelings	States or implies that the character is feeling romantic.	aroused, attracted, flirtatious, love, fond, desire, adore, admire	positive
trust	States or implies that the character believes in something/someone else.	trust	positive
shocked/surprised	States or implies that the character is stricken by something unexpected/unbelievable.	shocked, surprised, disbelief	neutral
attentive/focused/engaged	States or implies that the character is focusing on something or actively engaged in something, with a neutral valence.	attentive, concentrated, concerned, intense, serious	neutral
confused/unsure	States or implies that the character has difficulty understanding something or making a decision	confused, puzzled, uncertain	neutral
contemplating	States or implies that the character is in thought, ruminating, or is thinking about something hard. (Also code "1" if the response is the content of the character's thought.)	in thought, processing, trying to sort a problem out, overthinking	neutral
determined	States or implies that the character has made a firm decision and/or is prepared to do something.	determined, prepared, responsible, persuasive	neutral
neutral	States or implies that the character is not thinking/feeling anything special.	blank, neutral, numb	neutral
not surprised	States or implies that the character has expected a certain outcome and does not feel surprised about it.	not surprised, knew it	neutral

pretending	States or implies that the character is pretending to be thinking about/feeling something different from what he/she is actually thinking about/feeling.	faking, pretending, trying to act as if	neutral
angry/irritated	States or implies that the character is in a state of agitation, anger or aggressiveness, with a higher arousal.	agitated, angry, annoyed, fed up, hate, frustrated	negative
anxious/scared	States or implies that the character is in a state of apprehension, nervousness or fear. (If a negative valence is not implied, code as attentive/focused/engaged)	nervous, anxious, scared, worried, stressed	negative
dissent	States or implies that the character disagrees with or disproves something.	disagreeing, disapproving, confronting, dislike	negative
schadenfreude	States or implies that the character is feeling happy for another person being upset.	secretly happy, schadenfreude	negative
proud/arrogant	States or implies that the character is highly self-satisfied accompanied with contempt to others.	cocky, smug, sarcastic, dominant	negative
sinister intention	States or implies that the character is having some sort of evil intention.	deceptive, up to no good	negative
suspicious	States or implies that the character has doubt on something.	suspicious, doubting	negative
upset	States or implies that the character feels unhappy, with a lower arousal.	deflated, disappointed, distressed, sad, regret, jealous, guilty, overwhelmed	negative
awkward	States or implies that the character is uneasy or embarrassed not knowing what to say or what to do.	awkward, embarrassed, shame, uncomfortable, wants to leave	negative
uninterested/distracted	States or implies that the character does not feel engaged and/or has his/her thoughts drawn away.	bored, unamused, distracted	negative
n/a	Does not state or imply that the character is engaging in any mentalising	waiting, hungry, listening, tired	n/a

Appendix B. Calculation of alignment scores for studies 2 and 3

Example: Calculating the scores of participants Young1, Young2 and Young3 at the category level.

Same-group scoring

Step 1: Establishing the adjusted weights

Picture 1	**Category (j)**					
Participant	Happy	Confident	Neutral	Angry/irritated	Upset	Anxious/ scared
Young1	x_{11}=1	x_{12}=0	x_{13}=0	x_{14}=0	x_{15}=1	x_{16}=0
Young2	x_{21}=1	0	0	0	1	1
Young3	x_{31}=1	1	0	0	0	0
Total $(\sum_{i=1}^{n} x_{ij})$	**3**	**1**	**0**	**0**	**2**	**1**
Adjusted weight $=\frac{1}{n-1}\left(\sum_{i=1}^{n} x_{ij}-1\right)$	$\frac{3-1}{3-1}$	$\frac{1-1}{3-1}$	$\frac{0-1}{3-1}$	$\frac{0-1}{3-1}$	$\frac{2-1}{3-1}$	$\frac{1-1}{3-1}$

Step 2: Calculating cell weighted values = $\frac{1}{n-1}(x_{pj}\ \cdot(\sum_{i=1}^{n} x_{ij}-\ 1)) = (x_{pj})$**(Adjusted weight)**$_{j}$

Picture 1	**Category**						**Score =**
Participant	Happy	Confident	Neutral	Angry/ irritated	Upset	Anxious/ scared	**Average weighted value**
Young1	$(1)(\frac{3-1}{3-1})$ $= \mathbf{1}$	$(0)(\frac{1-1}{3-1})$ $= 0$	$(0)(\frac{0-1}{3-1})$ $= 0$	$(0)(\frac{0-1}{3-1})$ $= 0$	$(1)(\frac{2-1}{3-1})$ $= \frac{\mathbf{1}}{\mathbf{2}}$	$(0)(\frac{1-1}{3-1})$ $= 0$	$\frac{\mathbf{1}+\frac{\mathbf{1}}{\mathbf{2}}}{\mathbf{2}}$ $= \mathbf{0.75}$
Young2	$(1)(\frac{3-1}{3-1})$ $= \mathbf{1}$	$(0)(\frac{1-1}{3-1})$ $= 0$	$(0)(\frac{0-1}{3-1})$ $= 0$	$(0)(\frac{0-1}{3-1})$ $= 0$	$(1)(\frac{2-1}{3-1})$ $= \frac{\mathbf{1}}{\mathbf{2}}$	$(1)(\frac{1-1}{3-1})$ $= 0$	$\frac{\mathbf{1}+\frac{\mathbf{1}}{\mathbf{2}}}{\mathbf{2}}$ $= \mathbf{0.75}$

Young3	$(1)(\frac{3-1}{3-1})$ $= \mathbf{1}$	$(1)(\frac{1-1}{3-1})$ $= 0$	$(0)(\frac{0-1}{3-1})$ $= 0$	$(0)(\frac{0-1}{3-1})$ $= 0$	$(0)(\frac{2-1}{3-1})$ $= 0$	$(0)(\frac{1-1}{3-1})$ $= 0$	$\mathbf{\frac{1}{1} = 1}$

Alternative formula for cell weighted value:

$$\boldsymbol{\frac{1}{n-1}\left(x_{pj}\ \cdot\ \sum_{i=1}^{n} x_{ij} -\ x_{pj}\right) = \frac{1}{n-1}\left(x_{pj}\ \cdot\ total_j -\ x_{pj}\right)}$$

Picture 1	**Category**						**Score = Average weighted value**
Participant	Happy	Confident	Neutral	Angry/ irritated	Upset	Anxious/ scared	
Young1	$\frac{(1)(3)-1}{3-1}$ $= \mathbf{1}$	$\frac{(0)(1)-0}{3-1}$ $= 0$	$\frac{(0)(0)-0}{3-1}$ $= 0$	$\frac{(0)(0)-0}{3-1}$ $= 0$	$\frac{(1)(2)-1}{3-1}$ $= \mathbf{\frac{1}{2}}$	$\frac{(0)(1)-0}{3-1}$ $= 0$	$\mathbf{\frac{1+\frac{1}{2}}{2}}$ $= \mathbf{0.75}$
Young2	$\frac{(1)(3)-1}{3-1}$ $= \mathbf{1}$	$\frac{(0)(1)-0}{3-1}$ $= 0$	$\frac{(0)(0)-0}{3-1}$ $= 0$	$\frac{(0)(0)-0}{3-1}$ $= 0$	$\frac{(1)(2)-1}{3-1}$ $= \mathbf{\frac{1}{2}}$	$\frac{(1)(1)-1}{3-1}$ $= 0$	$\mathbf{\frac{1+\frac{1}{2}}{2}}$ $= \mathbf{0.75}$
Young3	$\frac{(1)(3)-1}{3-1}$ $= \mathbf{1}$	$\frac{(1)(1)-1}{3-1}$ $= 0$	$\frac{(0)(0)-0}{3-1}$ $= 0$	$\frac{(0)(0)-0}{3-1}$ $= 0$	$\frac{(0)(2)-0}{3-1}$ $= 0$	$\frac{(0)(1)-0}{3-1}$ $= 0$	$\mathbf{\frac{1}{1} = 1}$

Crossed-group scoring

Step 1: Establishing the weights

Picture 1	**Category**					
Participant	Happy	Confident	Neutral	Angry/irritated	Upset	Anxious/scared
Old1	y_{11}=1	y_{12}=1	y_{13}=0	y_{14}=0	y_{15}=0	y_{16}=0
Old2	y_{21}=0	0	1	0	0	0
Old3	y_{31}=0	0	0	1	0	1
Total $(\sum_{i=1}^{3} y_{i1})$	**1**	**1**	**1**	**1**	**0**	**1**
Weight $=\frac{1}{n}\sum_{i=1}^{n} y_{ij}$	$\mathbf{\frac{1}{3}}$	$\mathbf{\frac{1}{3}}$	$\mathbf{\frac{1}{3}}$	$\mathbf{\frac{1}{3}}$	**0**	$\mathbf{\frac{1}{3}}$

Step 2: **Calculating cell weighted values =** $\frac{1}{n}\left(x_{pj}\ \cdot\ \sum_{i=1}^{n} y_{ij}\right)$ = $(\boldsymbol{x_{pj}})$**(weight)$_j$**

Picture 1	**Category**						**Score = Average weighted value**
Participant	Happy	Confident	Neutral	Angry/ irritated	Upset	Anxious/ scared	
Young1	$(1)\left(\frac{1}{3}\right)$ $=\frac{1}{3}$	$(0)(\frac{1}{3})$ $=0$	$(0)(\frac{1}{3})$ $=0$	$(0)(\frac{1}{3})$ $=0$	(1)(0) =0	$(0)(\frac{1}{3})=0$	$\mathbf{\frac{\frac{1}{3}}{1}=0.67}$
Young2	$(1)\left(\frac{1}{3}\right)$ $=\frac{1}{3}$	$(0)(\frac{1}{3})$ $=0$	$(0)(\frac{1}{3})$ $=0$	$(0)(\frac{1}{3})$ $=0$	(1)(0) =0	$(1)\left(\frac{1}{3}\right)$ $=\frac{1}{3}$	$\mathbf{\frac{\frac{1}{3}+\frac{1}{3}}{2}}$ $\mathbf{=0.67}$
Young3	$(1)\left(\frac{1}{3}\right)$ $=\frac{1}{3}$	$(1)\left(\frac{1}{3}\right)$ $=\frac{1}{3}$	$(0)(\frac{1}{3})$ $=0$	$(0)(\frac{1}{3})$ $=0$	(0)(0) =0	$(0)(\frac{1}{3})=0$	$\mathbf{\frac{\frac{1}{3}+\frac{1}{3}}{2}}$ $\mathbf{=0.67}$

Summary

	Participant	Score (same-group)	Score (crossed-group)
Picture 1	Young1	0.75	0.67
	Young2	0.75	0.67
	Young3	1	0.67

Assuming that there is a total of three items and the table below shows the scores of participants Young1, Young2 and Young3 on the other two items.

	Participant	Score (same-group)	Score (crossed-group)
Picture 2	Young1	0.8	0.33
	Young2	0.6	0.67
	Young3	0.4	0.18
Picture 3	Young1	0.25	0.40
	Young2	0.75	0.60
	Young3	0.5	0.33

The final scores of the three participants are calculated by taking the average of their scores across the three items. These are the scores to be compared in the paired t-test.

Participant	Mean score (same-group)	Mean score (crossed-group)
Young1	$\frac{0.75+0.8+0.25}{3} = \mathbf{0.6}$	$\frac{0.67+0.33+0.4}{3} = \mathbf{0.47}$
Young2	$\frac{0.75+0.6+0.75}{3} = \mathbf{0.7}$	$\frac{0.67+0.67+0.6}{3} = \mathbf{0.65}$
Young3	$\frac{1+0.4+0.5}{3} = \mathbf{0.63}$	$\frac{0.67+0.18+0.33}{3} = \mathbf{0.39}$

Appendix C. List of context and interpretations (options) presented in Study 5a-c

Study 5a.

Picture	High-frequency context	Low-frequency context	Interpretations
P3	They are colleagues.	They are a couple.	He is interested in what she is saying. (High-frequency target)
			He is in love with his partner. (Low-frequency target)
			He is wondering whether this date is worth it.
			He is happy with his dinner companion and hoping she feels the same.
P4	They are on an elite training programme at work.	They are sisters.	She is feeling empathy for the other person.
			She is feeling upset about something which has happened. (Low-frequency target)
			She is quietly pleased with what is happening. (High-frequency target)
			She is angry at what has happened to the person next to her.
P5	They have recently met.	They have known each other for a long time.	He is feeling amused.
			He feels attracted to her. (High-frequency target)
			He is remembering something that happened between them earlier. (Low-frequency target)
			He is feeling relaxed.
P8	They've been told that an incident has occurred inside.	They've been told that they	She is scared. (High-frequency target)
			She is surprised at what she has just heard.

		aren’t allowed in.	She is very sad about something somebody has said
			She is very offended. (Low-frequency target)
P9	He just got a call from his daughter.	He just called his daughter out for dinner.	He is focused and concerned about what she is saying. (High-frequency target)
			He is worried about the news he is about to pass on. (Low-frequency target)
			He is angry and annoyed about her attitude.
			He is feeling disappointed with his daughter.
P10	They are on their way to a meeting.	They are on their way home from lunch.	She is feeling angry with him. (Low-frequency target)
			She is concentrating on an upcoming meeting.
			She is problem solving.
			She is determined and about to take on a challenge. (High-frequency target)

Study 5b.

Picture	High-frequency context	Low-frequency context	Interpretations
P3	The two colleagues are having dinner together after a work meeting.	The two are sharing a meal on their anniversary.	He is interested in what she is saying. (High-frequency target)
			He is in love with his partner. (Low-frequency target)
			He is wondering whether this date is worth it.
			He is happy with his dinner companion and hoping she feels the same.
P4	The colleagues are on a competitive training programme at work.	The sisters just came home together from a difficult family gathering.	She is feeling empathy for the other person.
			She is feeling upset about something which has happened. (Low-frequency target)
			She is quietly pleased with what is happening. (High-frequency target)
			She is angry at what has happened to the person next to her.
P5	The new colleagues just discovered shared preferences in food and music.	The colleagues just discovered they were middle school classmates.	He is feeling amused.
			He feels attracted to her. (High-frequency target)
			He is remembering something that happened between them earlier. (Low-frequency target)
			He is feeling relaxed.
P8	They’ve been told that an incident has occurred inside.	They’ve been told they aren’t welcome and won’t be allowed in.	She is scared. (High-frequency target)
			She is surprised at what she has just heard.
			She is very sad about something somebody has said.

			She is very offended. (Low-frequency target)
P9	He just got an emergency call from his daughter.	He just invited his daughter out to tell her his decision to divorce her mother.	He is focused and concerned about what she is saying. (High-frequency target)
			He is worried about the news he is about to pass on. (Low-frequency target)
			He is angry and annoyed about her attitude.
			He is feeling disappointed with his daughter.
P10	They are on the way to handling a difficult assignment.	The couple is on their way home from lunch.	She is feeling angry with him. (Low-frequency target)
			She is concentrating on an upcoming meeting.
			She is problem solving.
			She is determined and about to take on a challenge. (High-frequency target)

Study 5c.

Picture	High-frequency context	Low-frequency context	Interpretations
P3	The two colleagues are having dinner together after a work meeting.	The two are sharing a meal on their anniversary.	He is interested in what she is saying. (High-frequency target)
			He is in love with his partner. (Low-frequency target)
			He is wondering whether this date is worth it.
			He is happy with his dinner companion and hoping she feels the same.
P4	The colleagues are on a competitive training programme at work.	The sisters just came home together from a difficult family gathering.	She is feeling empathy for the other person.
			She is feeling upset about something which has happened. (Low-frequency target)
			She is quietly pleased with what is happening. (High-frequency target)
			She is angry at what has happened to the person next to her.
P5	The colleagues have found that they are both free after work today.	The colleagues have been friends since childhood.	He is feeling amused.
			He feels attracted to her. (High-frequency target)
			He is remembering something that happened between them earlier. (Low-frequency target)
			He is feeling relaxed.
P8	They’ve been told that it hasn’t yet been possible to contact their daughter.	They’ve been told that their reservation was cancelled.	She is feeling anxious. (High-frequency target)
			She is feeling shocked.
			She is very sad about something somebody has said.

			She is feeling hugely annoyed. (Low-frequency target)
P9	He just got an emergency call from his daughter.	He just invited his daughter out to tell her his decision to divorce her mother.	He is focused and concerned about what she is saying. (High-frequency target)
			He is worried about the news he is about to pass on. (Low-frequency target)
			He is angry and annoyed about her attitude.
			He is feeling disappointed with his daughter.
P10	They are on the way to handling a difficult assignment.	The couple is on their way home from lunch.	She is feeling angry with him. (Low-frequency target)
			She is concentrating on an upcoming meeting.
			She is problem solving.
			She is determined and about to take on a challenge. (High-frequency target)

Appendix D. Calculation of proxy of the probability of generating an interpretation in Context 1 for Study 8

Example:

Calculating the proxy of the target probability for participant 0001 for items 1 and 2.

Step 0: Raw data

Table A

Participant	**Item**	**Interpretation (response)**	**Context 1 rating**	**Context 2 rating**
0001	1	a_1	4	2
0001	1	b_1	3	3
0001	1	c_1	5	3
0001	1	d_1	4	4
0001	2	a_2	3	3
0001	2	b_2	4	2

Step 1: Recode the likelihood ratings in both contexts into 0 and 1 based on whether they fall below the scale mid-point (1-3) or exceeds the scale mid point (4-6)

Table A

Participant	**Item**	**Interpretation (response)**	**Context 1 rating**	**Context 2 rating**	**Context 1 rating (dummy)**	**Context 2 rating (dummy)**
0001	1	a_1	4	2	1	0
0001	1	b_1	3	3	0	0
0001	1	c_1	5	3	1	0
0001	1	d_1	4	4	1	1
0001	2	a_2	3	3	0	0
0001	2	b_2	4	2	1	0

Step 2: Set up a new table that lists out the possible combinations of the possible values of the two dummy variables for each item.

Table B

Participant	**Item**	**Context 1 rating (dummy)**	**Context 2 rating (dummy)**
0001	1	1	1
0001	1	1	0
0001	1	0	1
0001	1	0	0
0001	2	1	1
0001	2	1	0
0001	2	0	1
0001	2	0	0

Step 3: Tally the number of responses that match each possible combination of the possible values of the two dummy variables for each item based on the table in Step 1.

Table B

Participant	**Item**	**Context 1 rating (dummy)**	**Context 2 rating (dummy)**	**Number of matching responses**
0001	1	1	1	1
0001	1	1	0	2
0001	1	0	1	0
0001	1	0	0	1
0001	2	1	1	0
0001	2	1	0	1
0001	2	0	1	0
0001	2	0	0	1

Step 4: Count each participant's total number of interpretations generated for each item.

Table B

Participant	**Item**	**Context 1 rating (dummy)**	**Context 2 rating (dummy)**	**Number of matching responses**	**Number of total responses**
0001	1	1	1	1	4
0001	1	1	0	2	4
0001	1	0	1	0	4
0001	1	0	0	1	4
0001	2	1	1	0	2
0001	2	1	0	1	2
0001	2	0	1	0	2
0001	2	0	0	1	2

Step 5: Calculate the proportion of responses matching each possible combination of dummy variable values as $\frac{Number\ of\ matching\ responses}{Number\ of\ total\ responses}$.

Table B

Participant	**Item**	**Context 1 rating (dummy)**	**Context 2 rating (dummy)**	**Number of matching responses**	**Number of total responses**	**Proportion of matching responses**
0001	1	1	1	1	4	1/4 = 0.25
0001	1	1	0	2	4	2/4 = 0.5
0001	1	0	1	0	4	0
0001	1	0	0	1	4	1/4 = 0.25
0001	2	1	1	0	2	0
0001	2	1	0	1	2	1/2 = 0.5
0001	2	0	1	0	2	0
0001	2	0	0	1	2	1/2 = 0.5

The highlighted column (Proportion of matching responses) is the proxy of the target probability and is used as the dependent variable in the linear mixed model, as explained in Figure 6.1.