Compensation and Theory of Mind: An investigation into the counter compensatory value of decreasing the answer and stimuli presentation time of theory of mind assessments

A thesis presented in partial fulfilment of the requirement for the degree of

Master of Science

in

Psychology

at Massey University, Manawatū, New Zealand

Liam Joshua Allen

2020

Abstract

Over the last few decades there has been a huge growth of interest in the concept of theory of mind (ToM). However, recent research suggests that many ToM assessment have a range of psychometric issues, one recently discerned and highly problematic issue of which is that many individuals with significant ToM deficits and real-world social difficulties are able successfully pass ToM assessments through the use of compensatory strategies, invalidating the assessment(Livingston et al., 2019b). One way to potentially fix this issue and improve the assessments psychometric properties may lie in the addition of increased answer or stimuli presentation time constraints. To test this theory we designed an experiment based around three manipulated versions of the Reading the Mind in the Eyes Test (RMET-R). These three variations forming the main IV of the experiment and consisting of a 'long' variation with a 20 sec answer time, a 'short' variation with a 5 sec answer time, and a 'occluded' variation with a 0.5 second stimuli presentation time limit, and a 5 sec answer time limit. As we lacked access to a clinical sample we chose to assess if our three experimental conditions moderated the relationship between RMET-R scores and two other theoretically related measures. 381 participants were recruited from Prolific academic to conduct an online experiment in which they randomly completed one of our three RMET-R variations and a series of other measures including the TEQ and MentS. Contrary to as we hypothesised the results analyses indicated that shortening the available answer time of the assessment significantly increased the difficulty of the RMET-R, but not its validity or reliability. And that shortening the presentation time of the RMET-R's stimuli had no effect on the test's difficulty, validity or reliability. However it is quite possible that our sample simply lacked the compensating individuals that were required to test our theory, and that our results simply indicate the effect of our experimental constraints on neurotypical individuals.

Unexpectedly this pattern of results suggests that our experimental constraints could alternatively be quite useful in increasing the ecological validity of ToM assessments.

Acknowledgements

I would like to thank and acknowledge the contributions of a number of people that have been indispensable over this past year and contributed greatly to my academic endeavours. The first of which is a huge thank you to my supervisor Michael Philipp and co-supervisor Stephen Hill who have both been a great font of knowledge and guidance throughout this process. I'd also like to thank Malcolm Loudon who without his technical know-how none of this would have been possible. A very special thank you goes to my mother Sonya as well who has always been a great source of wisdom, love, support, and determination. I would also like to give a shout out to my younger brother Taylor, who has been a great help throughout this year, and a brilliant sounding board. A very special thank you to my father Trevor as well, who's been a great source of support and wisdom. A special hello to my youngest sibling ashe as well, who I hope is going well.

I would also like to thank all my other friends and family that I haven't mentioned by name, and all the participants of this study who have literally gotten me this far, thank you for everything.

Table of Contents

List of Tables

List of Figures

CHAPTER 1

Introduction

"interest in theory of mind began as the province of handful of researchers, now a legion of researchers study theory of mind across the world" Wellman 2018b.

Over the last couple of decades interest in Theory of Mind (ToM) has grown exponentially, with a great swathe of research performed on the topic. The concept of ToM refers to an individual's social-cognitive ability to identify, understand, and attribute mental states of other individuals, allowing them to predict others behaviours and react accordingly. Interest in ToM has expanded far beyond its initial roots in the study of animal and infant metacognition in the late 1970s, to the point that it is a major topic of interest amongst several sub-fields of psychology today and referenced by a great deal more. Wellman (2018b) noting that numerous researchers now study ToM across the world, whose research over the years has led to the generation of a great deal of established practices, findings, and tools. This has led to the creation of a particularly large and variable array of assessments designed to measure individual's ToM abilities (Wellman, 2018a). At the current time much of this field of research is in a state of flux. Recent research has challenged a number of previously well-established practices and theories and directing researchers down a range of newly discerned pathways. Some of the earliest researchers of ToM note that one key takeaway from the decades of research that has been performed is that there is still much left to uncover (Wellman, 2018b).

This current state extends to the array of assessments that have been developed, with recent research proposing most assessments in common use today have multiple psychometric issues. One recently discerned and highly problematic issue of which is that many assessments are not able to validly assess compensating individuals (Livingston et al.,

2019b). Recent research suggests many individuals with significant underlying ToM deficits and real-world social difficulties are able to use compensatory techniques or processes to successfully boost their ToM abilities and camouflage their underlying deficits. Theoretically these techniques enable such individuals to successfully pass such assessments, though likely at an increased cost to the time and mental effort required. Which invalidates their assessment results and raises a number of qualms about the supposed validity of the assessment. Recent research also suggests that the use of such compensatory mechanisms is fairly common amongst clinical populations with well-known poor ToM skills, making this a larger issue than originally thought (Livingston & Happé, 2017; Hull et al., 2019).

One potential fix for this problem may lie in the addition of increased answer and stimulus time constraints to existing ToM assessment questions. These additional constraints decrease the amount of time participants have to answer each of the assessment's questions and limits their view of the stimuli the questions are focussed on. This theoretically denying compensating individuals the extra time they require to utilise their compensatory strategies and/or processes. And hypothetically increases the difficulty and cognitive demands of the assessment to the point that compensating individuals' more fragile abilities begin to faulter.

The addition of multiple time or stimulus constraints within a single assessment may also allow for new avenues of analysis within older assessments. This potentially allows assessors to gradually increase the difficulty of the assessment, enabling the assessment for abnormal scoring differences between different time conditions. Assuming that these additional constraints or the splitting of the assessment do not unduly affect the reliability of the assessment. This thesis is built around the investigation of the potential value these two additional constraints. Despite their simplicity and use in other cognitive assessments these two methods have not been widely explored in ToM assessment; with no research currently

having been performed on the value of multiple answer times in a single ToM assessment, or the addition of increased stimulus limitations.

This chapter provides a review of the current conceptualisation of ToM, briefly touching on its history, importance, neurological basis, similarities to other constructs and its current competing conceptualisations. It also discusses the current methods of assessing ToM, including the most commonly used assessments and several of the ongoing problems within them. Furthermore, a brief look at some current research on compensation will be included, and what exactly this means for the supposed validity of many ToM assessments.

Following this review the hypothetical value of the addition of increased answer and stimulus time constraints will also be described in more detail, along with its theoretical value and some research which supports said value. Finally, the main aims and hypotheses of this study will be presented along with some information on a couple of secondary exploratory aims. This literature review aims to provide a solid background to ToM and its assessment and its current problems, in order to help explain and highlight the rationale for the conduction of this study.

What is Theory of Mind?

There is strong support for the claim that ToM is absolutely central to everyday human life and has played a huge role in the development of humanity as a species (Brüne, & Brüne-Cohrs, 2006). Theory of Mind refers to an individual's ability to infer and reason about other individuals' mental states which allows them to make predictions about other people's beliefs, dispositions and intentions. This in turn provides the ability to understand and predict other individuals' behaviours, which is crucial for successful social interaction and communication (Brüne, & Brüne-Cohrs, 2006). Theory of Mind in effect literally forms

the ability to create a theory about another's mind and forming a core component of an individual's social skills.

These intuitive and predictive skills provided by ToM are vital for the successful navigation of everyday social lives, playing a vital role in enabling social co-ordination and social fluidity (Canty, 2016). Almost all individuals required to engage in a diverse array of social interactions with a diverse series of goals on a daily basis (Brüne, & Brüne-Cohrs, 2006). Success in which is often of great importance and can prove vital for success in everyday life. These skills also play a central role in the development and acquisition of language, social norms and shared habits (Veissière et al., 2019). With ToM forming a vital prerequisite for social learning.

Individuals with ToM deficits typically have significantly poorer social capabilities and levels of social success especially when under increased pressure or engaging in more complex social behaviours such as persuading, deceiving, or flirting with others (Byom & Mutlu, 2013). Such individuals find it much more difficult to infer other individuals' intentions, appropriately reciprocate, and determine how their behaviour will affect others (Banerjee et al., 2011; Brüne, & Brüne-Cohrs, 2006). Higher levels of impairment have been found to be correlate with poorer social functioning, mental health, and lower quality of life (Milders et al., 2003; Livingston & Happé, 2017); as well as decreased peer popularity (Slaughter et al., 2015).

Conditions linked to ToM deficits

Due to ToM's critical role in social interaction a large amount of research has been performed investigating its potential role in a large variety of neuropsychiatric conditions and clinical populations that have well known social interaction difficulties. ToM deficits have been suggested to play a role in a sizable number of conditions, the most well-known and

researched is individuals with Autism Spectrum Disorders (ASD). The presence of ToM deficits in individuals with ASD is widely acknowledged and supported by extensive research findings over the last few decades (Baron-Cohen, 2000; Jones et al., 2018; Baron-Cohen, 1991b). Considerable research has also been performed which supports a link between schizophrenia and impaired ToM, though the exact nature of this impairment is less well defined then in ASD and an area of contention between researchers (Brüne, 2005; Sprong, Schothorst et al., 2007; Dimopoulou et al., 2017). A variety of research has also been performed in adults with several different forms of brain damage, with ToM deficits discovered in individuals with right hemisphere (Hamilton et al., 2017; Pluta et al., 2017), frontal lobe (Stone et al., 1998), and amygdala damage (Castelli et al., 2000).

A widespread array of more isolated research has also been performed searching for possible ToM deficits in a variety of other conditions. Impaired abilities have also been discovered in individuals with; epilepsy (Stewart et al., 2016), a variety of Personality disorders (Fossati et al., 2017; Petersen et al., 2016), bipolar disorder (Shamay-Tsoory et al., 2009; Bora et al., 2016), depression (Inoue et al., 2006; Bora, & Berk, 2016), eating disorders (Laghi et al., 2014), alcoholism (Bosco et al., 2013), Alzheimer's disease and other dementia disorders (Gregory et al., 2002; Kipps & Hodges, 2006), Huntington's disease (Eddy et al., 2012), language disorders (Spanoudis, 2016), and amongst individuals that have been emotionally or physically abused (Rnic et al., 2018). Deficits in ToM have also been found amongst prelingually deaf children, though this appears to be linked to language acquisition rather than ToM itself (Peterson, & Siegal, 1995). These deficits typically disappear by adulthood depending on when deaf children learn sign language, with late signers having more pronounced delays and poorer prognoses then early signers (Peterson et al., 2016a; O'Reilly et al., 2014).

Conception and Short history of ToM

The concept of ToM was originally proposed by the primatologists Premack and Woodruff (1978) in their seminal article: 'Does the chimpanzee have a theory of mind?'. The article introduced the idea that chimpanzees can infer the mental states of other chimpanzees and human actors. The phrase "Theory of Mind" concocted by the authors of this study to describe this ability. This construct was quickly adopted to explain this ability in humans, fitting in with a number of other emerging theories at the time in the fields of developmental psychology and metacognition. With the construct beating a number of other similar competing conceptualisations (Wellman, 2014; Wellman, 2018b). Largescale study of ToM didn't start until it was linked with autistic spectrum disorder by Baron-Cohen, Leslie, and Frith's (1985) article "Does the autistic child have a "theory of mind"?". This seminal study purposed that the behavioural symptoms of autistic spectrum disorder was caused by defects in such individuals ToM abilities. This theory spearheaded a vast amount of successive research and brought ToM into the public eye.

This interest has turned ToM into one of the most widely known and researched conceptions of social cognition (Leudar et al., 2004); as such a great deal of research into ToM's development (Wellman et al., 2001; Wellman, 2018b), theoretical processes (Brüne, & Brüne-Cohrs, 2006), neurological co-ordinates (Molenberghs et al., 2016; Schurz et al., 2014), relationship to other concepts (Milligan et al., 2007; Ahmed & Miller, 2011), assessment (Canty, 2016; Bora et al., 2009), and conditions that are associated with deficits (Sprong et al., 2007; Brewer et al., 2017) have been performed. ToM today is researched and referenced by a range of specialists; with developmental, clinical, experimental, social, and cognitive psychologists, neuropsychologists, neurologists, geneticists, educational researchers, philosophers, anthropologists, and primatologists, along with legal, and religious scholars all involved in its study (Wellman, 2018b).

Development of ToM

Theory of Mind abilities are not unique to primates, with multiple species such as ravens (Bugnyar et al., 2016) and dogs (Maginnity & Grace, 2014) found to possess some form of basic ToM abilities. In primates (Krupenye et al., 2016) and humans these abilities are much more advanced (Brüne, & Brüne-Cohrs, 2006). In addition to these much more advanced abilities humans also have a particularly strong intuitive desire to attribute agency to external subjects, to the point that the incorrect ascribing of intentions and behavioural traits to inanimate objects is fairly common (Brüne, & Brüne-Cohrs, 2006). This incorrect attribution of agency is somewhat more common amongst younger individuals but can be found throughout an individual's lifespans. Examples of which range from the personification of comfort objects by young children (Taylor et al., 2013), to adults' personification of financial markets and monetary based risks (Bossaerts et al., 2018). This common misrepresentation of agency pointing to just how important and how strongly ToM abilities are interwoven into humankind.

The development of more advanced ToM skills is intertwined with the development of humanity from an evolutionary perspective. The development of more advanced ToM abilities playing a key role in the development of the earliest pre-human societies as it enabled for successful functioning of larger and more complex social groups. This increase in group size resulted in a number of significant benefits for early humans including increased group security and individual safety, the capability to collectively solve problems, and a greater resource collection and storage abilities (Brüne, & Brüne-Cohrs, 2006). The development of ToM has also been noted to have played a central role in the development of language and culture and is partially responsible for the much slower reproduction and maturation period of humans, compared to other animals (Dunbar, 2003; Joffe, 1997). Brüne, & Brüne-Cohrs (2006) provide a more in-depth discussion of this topic.

Theory of Mind appears to be an innate potential ability for neurotypical individuals, its ontogeny not much different from the development of other cognitive functions and abilities (Brüne, & Brüne-Cohrs, 2006). An individual's ToM faculties take years to develop and require extensive social experience to fully mature. Individuals that are denied such experience during critical periods of development fail to fully develop their potential abilities and experience severe difficulties in later life (Peterson et al., 2016). A number of examples of which can be seen in cases of extreme childhood abuse and confinement. The development of ToM typically follows a fairly predictable pattern, slowly developing in tandem with other language and social skills over childhood (Milligan et al., 2007). Theory of Mind having a complicated relationship with language, with neither of these two-phenomenon exiting independently of each other.

The first sign or milestone of ToM development is the emergence of joint attention which is the ability to follow others' verbal and non-verbal cues to jointly focus on a specific object around the age of 12 months (Baron-Cohen, 1991a). This is then followed by the development of pretend play, at around 18-24 months (Brüne, & Brüne-Cohrs, 2006) and the ability to pass first order false belief tasks at age 3-4, which is the ability to understand that other individuals can hold incorrect beliefs (Hollebrandse, van Hout & Hendriks, 2014). Between the age of 5-7 most children gain the ability to pass second order false-belief tasks, which require the ability to understand that other individuals can hold false beliefs about other individuals' beliefs (Blackburn et al., 2015). Children also begin to understand metaphors and irony and start to reliably distinguish jokes from lies around this stage, with these abilities typically appearing from the age of around 6-7 (Brüne, & Brüne-Cohrs, 2006). The last major milestone that has been investigated in any great detail in the research literature is the ability to recognise and understand 'faux pas', which is when an individual says something socially incorrect or rude without realising it. The ability to reliably recognise

a ‘faux pas’ typically develops between the age of 9-11, though there is a higher degree of variation with this ability than other milestones (Baron-Cohen et al., 1999).

This pattern of milestones is well cemented with multiple research findings supporting this pattern of development, though the exact ages at which milestones are reached and their nature is still a matter of contention between different researchers (Kuntoro et al., 2017; Wellman, 2018a). A full description of this research and these developmental milestones is beyond the scope of this review. Wellman (2018a) and Brüne and Brüne-Cohrs (2006) both provide a good overview of the topic.

There are also a number of factors that have the ability to influence individual ToM development. The presence of multiple siblings and growing up in a bi-lingual environment for instance have both been found to have a positive effect on ToM development across multiple studies (McAlister & Peterson, 2007; Kovács, 2009). Several studies have also noted differences in the development of ToM between collectivist (e.g., Asian/Middle east) and individualist (e.g., European/American) cultures, though the results of such research are inconsistent. Some research supports the presence of significant differences between collectivist and individualist cultures (Shahaeian et al., 2011; Wellman, et al., 2006), while others support a universal developmental trajectory that exists regardless of cultural background (Liu et al., 2008; Kuntoro et al., 2017). Though there is some dispute the majority of research also supports a decline in ToM function in old age. Though one that is at a lesser degree compare to other cognitive functions and may be solely related to general cognitive decline (Reiter et al., 2017; Phillips et al., 2002).

Current theoretical perspectives

When it comes to the theoretical conceptualisation of ToM several divergent conceptualisations exist, with its exact framework a matter of ongoing debate (Brüne, &

Brüne-Cohrs, 2006). There is significant disagreement between researchers on even the base structure of the construct, with ongoing argument over whether ToM processing is a single or multidimensional construct with several single and multiple construct models existing (Canty, 2016; Carruthers, 2016). Proponents of multidimensional theories arguing that no single system account could allow for both efficient and flexible processing of social input (Apperly, 2013). While proponents of single dimensional setting argue that a single system that works in concert with alternative cognitive systems, such as executive control, empathy, and memory, makes more sense (Westra, 2017; Christensen, & Michael, 2016).

The most prominent current conceptualisation of ToM is that it is made up of two components: a social-cognitive/cognitive subprocess and a social-perceptive/affective subprocess. This theory was originally put forth by Tager-Flusberg and Sullivan (2000) and mirrors a similar theoretical conceptualisation of empathy (Dvash, & Shamay-Tsoory, 2014). Tager-Flusberg and Sullivan (2000) proposing that an individual's ToM abilities actually involve two distinct components: a "Cognitive ToM" process which involves individuals' wider shared world knowledge and factual/conceptual-representational inferring abilities and a "Affective ToM" process which refers to individuals' more rapid emotional and physical cue-based inferring abilities. Cognitive ToM abilities take longer to function and draw upon other cognitive functions such as executive functions and language, while affective ToM relies on a faster more automatic system.

The majority of ToM assessments are also commonly divided in this manner, such assessments either assessing individuals affective or cognitive ToM abilities with only a few designed to assess both. These two types of assessments are typically found to poorly correlate with each other, which adds further credence to the theory (Ahmed & Miller, 2011; Baron-Cohen et al., 1999). Findings from a variety of neuroimaging studies also add support for this theory, with different neural networks found to be reliably engaged between these

two differing types of ToM tasks (Molenberghs et al., 2016; Shamay-Tsoory et al., 2007). And damage to specific brain regions found to be related to selective affective/cognitive ToM deficits in a few studies (Shamay-Tsoory et al., 2006). Despite the wide support of this theory a few researchers have suggested that this division doesn't go quite far enough and have proposed that a model with further divisions may be more accurate. These researchers pointing to the wide lack of correlations between ToM assessments even within the supposed affective/cognitive divides to back up their claims (Warnell, & Redcay, 2019).

Several other similar multi-dimensional theories have also been proposed by a variety of researchers, which are commonly mixed up in the research literature (Frith, & Frith, 2008; Apperly, & Butterfill, 2009; Bohl, & van den Bos, 2012). For example, the two systems account of ToM also suggests that humans possess two processes that work in parallel with one another to enable ToM abilities. Which consists of an earlier developed, automatic "implicit" system and a slower, more effortful, and flexible "explicit" one, which slowly develops over childhood (Low, & Perner, 2012). Individuals relying on the implicit system when they need to rapidly anticipate others, and the explicit system when they need to more carefully reflect on specific individuals' beliefs.

Despite the wide popularity of these multidimensional theories there are a number of opponents, who argue that these dual theories are not supported by existing evidence. These opponents claim that ToM processing is fast, flexible, and context sensitive, regardless of what kind of social information is processed. (Carruthers, 2016; Westra, 2017). Although these single system models are less popular today, there have been several prominent models which have been developed; which include the ToM module (Gerrans, 2002), theory-theory perspective (Perner, 1991), and simulation theory models (Gallese, & Goldman, 1998).

Out of the three the ToM module is the simplest theory and proposes that all ToM processing is conducted by a single dissociable neural network that inflexibly matures over childhood (Gerrans, 2002). The theory-theory perspective is not solely a theory of ToM but also provides an explanation for it. The theory-theory perspective attributes ToM abilities to tactical theoretical reasoning about known causal states. An individual's ToM abilities based on their ability to relate the reasons behind others behaviour and likely mental states based on their prior knowledge about their own inner states, motivations, and behaviours (Perner, 1991). The theory suggests that all individuals effectively act as naïve scientists, with ToM driven in all individuals by a continually evolving scientifically derived algorithm based on one's past experiences and observations (Brüne, 2005). The simulation theory meanwhile proposes that the ability to infer others' mental states is driven by the intuitive ability to cognitively simulate others' attributed mental states in one's own mind (Brüne, 2005). This theory proposes that ToM is closely related to empathy. And is based on research that suggests humans have a highly developed system of mirror neurons that are activated when they observe certain socially related movements and facial expressions in others (Canty, 2016; Gallese, & Goldman, 1998).

Over the last two decades the development of neuroimaging and brain mapping techniques has also allowed for another means to investigate the processes behind ToM, allowing for the localization of brain regions involved (McPartland, 2019). Such neurological research has reliably identified a core ToM network that involves the reliable engagement of the anterior dorsal medial prefrontal cortex (mPFC) and bilateral temporoparietal junction (TBJ) regardless of the type of ToM task (Molenberghs et al., 2016; Wade et al., 2018). A sizable amount of neuroimaging research has also been found which supports a multi-dimensional conceptualisation of ToM. Several additional regions have been found to reliably

engaged depending on the type of task in addition to brain regions in the core ToM network (Molenberghs et al., 2016).

Despite the intensive scale of research done on this topic, the true structure of ToM remains a matter of conjecture. As to what the best framework available is remains a matter of contentious debate with no clear answer available. A number of researchers suggest the best answer available at this time is made up of a mix of these aforementioned conceptualisations (Schaafsma et al., 2015; Apperly, 2012).

Similarities to other concepts

One problematic and widely criticised issue with the current state of social cognition research is that the precise definitions and boundaries of the concepts that fall under the umbrella of social cognition, which include ToM, are vaguely specified and poorly defined (Adolphs, 2001; Schaafsma et al., 2015; Happé et al., 2017). This is due to a poor level of consensus amongst researchers about the exact definitions of the concepts that fall under the umbrella of social cognition. For example, there remains substantial disagreement over even how many processes social cognition is made up of. Fiske, & Taylor (2013) argue that there are 14 distinct processes while Happé, & Frith (2014) suggest the presence of 10 and Green, Horan, & Lee, (2015) argue that there are 4. This has led to a high level of chaos with substantial overlap between many concepts and competing definitions of the same terms existing.

Many researchers use the same terms to refer to different concepts. The terms ToM, mentalizing, mindreading, social intelligence, and mind perception, are used by authors interchangeably to refer to the same overarching ToM concept; while at the same time others believe that there are more concrete boundaries between these separate concepts (Schaafsma et al., 2015). As an example; a subgroup of researchers exists that categorises mentalization

as a separate construct that overlaps with ToM, that additionally involves self-reflection of one's own and others' mental states; while other individuals use it as an alternative term to refer to ToM (Fonagy, & Allison, 2014; Dimitrijević et al., 2018). This issue also extends to ToM itself. The cognitive ability to understand one's own state of mind is described as part of ToM by some authors (Brüne, & Brüne-Cohrs, 2006) but described as a separate construct by others (Happé et al., 2017).

Even amongst some of the more broadly agreed upon concepts there is substantial overlap present. For example, affective ToM, cognitive empathy, and emotion recognition are so similarily defined as to be nearly interchangeable but are commonly noted to be separate constructs. Neuroimaging research supports the position that at least the broader concepts of social cognition, such as empathy and ToM, do refer to separate processes: with empathy (Kanske et al., 2015; Dvash, & Shamay-Tsoory, 2014), ToM (Schaafsma et al., 2015; Molenberghs et al., 2016), and emotional recognition (Murphy, Nimmo-Smith, & Lawrence, 2003) relying on overlapping but distinct neural networks.

The concept of ToM is also similar to a number of concepts developed outside the field of psychology. The concept is very similar to the philosophically based concept of Folk psychology (Ratcliffe, 2006) and the sociological theory of role-taking, or social perspective taking (Enright & Lapsley, 1980). These concepts were both derived before the conceptualization of ToM and have seen far less interest. Folk Psychology, for instance, is rarely referred to outside of philosophy today and despite seeing some interest between the 60's – 80's the concept of role-taking has fallen out of favour with the rise of ToM (Enright, & Lapsley, 1980).

The chaotic nature of these fields of research make it difficult for even seasoned experts to accurately navigate (Happé et al., 2017) and can be extremely confusing and

offputtingly for newcomers (Schaafsma et al., 2015). The field itself in desperate need of more concrete boundaries. Happé et al. (2017) provide a useful overview of this problem and offer some suggestions to help better organise the field.

ToM Assessments

The great amount of research that has been conducted on ToM has led to the development of a vast and diverse array of measures designed to assess individuals' ToM abilities. At least a hundred purposefully developed assessments that can be categorised as ToM assessments easily existing. Along with a score of alternative variations of the more commonly used assessments. With a particularly high number of measures designed for the assessment of advanced ToM abilities in individuals beyond late childhood.

This review will mainly focus upon this group of so called 'advanced' assessments as the more basic 'simple' assessments designed for infants and younger children between the ages of 0-8 have been found to be easily completed by older individuals with known deficits. The level of simplicity of such assessments make it unlikely that compensation would play a large role in their successful completion. Which lowers the potential value of the additional time constraints that make up the area of focus of this thesis. Regardless simple-ToM assessment in younger individuals will still be briefly gone over, to present a more comprehensive review of how ToM assessments have been developed over time.

As much of the early research on ToM focused on assessing the development of ToM in early childhood or autistic children this initially led to a deficit of more advanced measures (Valle et al., 2015). The first 'simple' ToM assessments began to be developed around the mid-1980's (Baron-Cohen et al., 1985), while the first 'advanced' measures only started to be developed around a decade later (Happé, 1994). It is worth noting that although they were not designed as ToM assessments the field of role-taking developed a number of measures that

could easily be classified or considered as at least precursors to ToM assessments (Feffer & Gourevitch, 1960; Chandler, 1973). A sizable number of assessments were developed to assess role-taking in children and adolescents between the 60s and 70s, which have fallen out of common knowledge (Enright, & Lapsley, 1980).

'Simple' ToM assessments

Compared to the assessment of more advanced ToM abilities amongst older individuals the assessment of ToM abilities in younger individuals between the ages of 0-8 is far more standardised (Apperly, 2012). The majority of 'simple' ToM assessments either assess first or second-order false belief abilities, such as the 'Sally–Anne' test designed by Baron-Cohen et al. (1985),or appearance-reality distinction abilities; where children are assessed on their ability to divorce what they know from what another individual will perceive, such as in the smarties test developed by Perner et al. (1987). A rare few more advanced assessments for children have been developed more recently. These include a couple of battery assessments for children such as the ToM Scale developed by Wellman and Liu (2004), which allow for assessment of multiple aspects of ToM to generate a clearer picture of a child's state of development. A variety of extremely simple, non-verbal assessments based on assessing an infant's gaze have seen a high level of interest over the last few years. This technique remaining a hot topic of interest today despite its dubious validity (Dörrenberg et al., 2018). However, these simpler assessments are useless for assessing older individuals (Baron-Cohen et al., 1999), with such assessments almost universally falling prey to ceiling effects when used with older individuals with all but the most severe ToM deficits (Bosco et al., 2014). This factor necessitated and led to the development of more advanced assessments in the mid 90's, when researchers become more interested in assessing ToM in older children and adults (Happé, 1994).

'Advanced' ToM assessments

Despite the shorter development period of advanced assessments, there has been a much greater number and variety of advanced assessments created. The last 20 years have seen a great increase of interest in assessing ToM among older populations, which has led to the development of a comprehensive selection of assessments designed to assess higher level ToM capabilities (Kaland et al., 2008). These assessments utilise a broad range of methodological designs and have been developed to be utilized in a range of complex contexts.

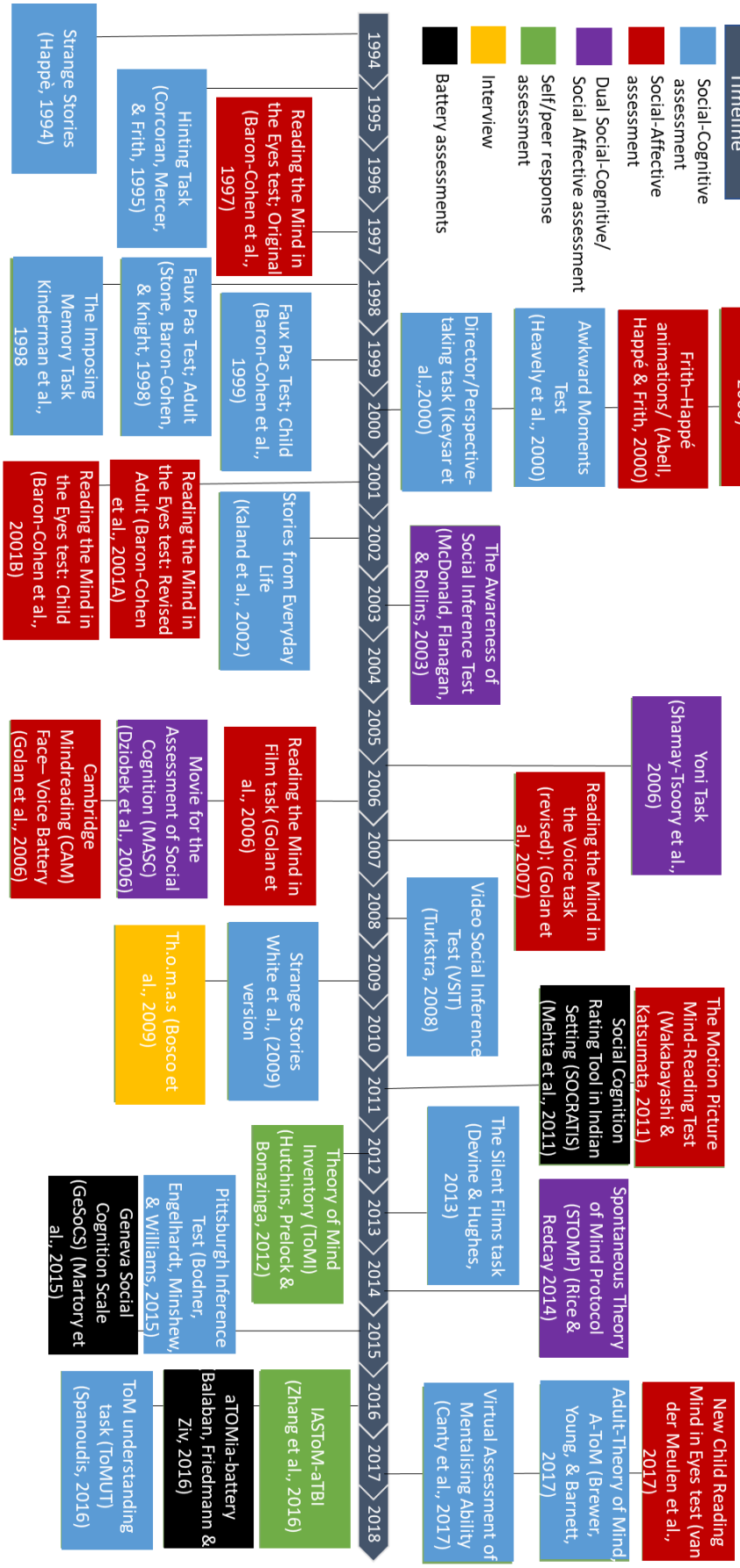
In contrast to assessments of other social-cognitive concepts such as empathy which are mainly focused around subjective self-answer-based designs, the vast majority of advanced ToM assessments are objective in nature. Most ToM assessments based on the presentation of some form of social stimuli which the assessments participants are then questioned about (Turner & Felisberti, 2017). The stimuli used in ToM assessments has become increasingly technologically advanced, older assessments typically using a series of pictures or written stories, while newer assessments commonly use video-based stimuli with even some forays into virtual-reality technology (Canty et al., 2017). Only a couple of subjective self-answer and peer review assessments such as the Theory of Mind Inventory (McDonald et al., 2003) and interview-based ToM assessments such as the Th.o.m.a.s (Bosco et al., 2009) exist.

As previously noted, many advanced ToM assessments are commonly divided between social-cognitive and social-affective assessments by researchers, following the dimensions of ToM proposed by Tager-Flusberg and Sullivan (2000). Social affective assessments such as the Frith–Happé animations (Abell et al., 2000) and Cambridge Mindreading (CAM) Face– Voice Battery (Golan et al., 2006a) focus on assessing rapid

judgment-based ToM abilities based on physical and emotional stimuli. While social cognitive assessments such as the Director/Perspective-taking task (Keysar et al., 2000) and Theory of Mind Understanding Task (ToMUT) (Spanoudis, 2016) focus on assessing individuals' abilities to assess the wider cognitive web of a number of other stimulated individuals. However, the obvious problems and limitations in only assessing a single aspect of ToM have become increasingly apparent to the wider scientific community over recent years (Warnell, & Redcay, 2019; Martory et al., 2015). This has led to an increase in the use of paired cognitive and affective ToM assessments in research studies. And has encouraged the creation of a number of dual social-cognitive/social-affective assessments such as the Movie for the Assessment of Social Cognition (MASC) (Dziobek et al., 2006), and dual cognitive/affective battery assessments like the Geneva Social Cognition Scale (GeSoCS) (Martory et al., 2015) to be developed. Figure 1 below displays the many of the more commonly used advanced assessments along with their original publication, and type of assessment.

Figure 1

Commonly used advanced Theory of Mind Assessments



Although many of the more recently developed assessments have been purposefully designed to improve upon and fix issues with earlier assessments, these older so called 'classical' assessments remain more widely used (Livingston et al., 2019a). However, this slow adaption of newly created assessments is not a recent problem, with multiple authors accusing this issue of impeding the progress of ToM research (Green, et al., 2005; Livingston et al., 2019a).

Over the last decade the most commonly used assessments have consisted of a variety of these 'classical' assessments. The most commonly used by a significant degree is the Reading the Mind in the Eyes Test – Revised (RMET-R) developed by Baron-Cohen et al. (2001a). Several other classical assessments, such as variations of the Short Stories test (Happé, 1994; White et al., 2009) and the adult (Stone et al., 1998) and child (Baron-Cohen et al., 1999) Faux pas tests follow in popularity.

A number of video-based assessments such as the Movie for the Assessment of Social Cognition (MASC) (Dziobek et al., 2006) and the inconsistently named Frith–Happé animations/Animated triangles task (Abell et al., 2000; White et al., 2011; Klein et al., 2009) have also seen some increased attention over the decade. The increased trend of computerised and video-based designs has also led to many of the older assessments receiving updates, with many of the most popular classical assessments adapted from pen and paper to computerised designs in recent research.

Exemplar ToM assessment: The Reading the Mind in the Eyes Test – Revised (RMET-R)

The RMET-R is an affective measure of ToM and revolves around the assessment of individuals' capabilities to determine mental states from photographs of actors' eyes (Baron-Cohen et al., 2001a). The RMET-R is the second iteration of the assessment and was designed to assess

mild ToM deficits in adult populations. The measure was originally created with autistic individuals in mind, though the assessment is not diagnostic in nature. The assessment was initially validated by its ability to successfully discriminate individuals with ASD and against a measure of autistic traits: the Autistic Spectrum Questionnaire (AQ) (Baron-Cohen et al., 2001A). An easier child version of the measure designed for younger individuals also exists, which was based on the same design and developed around the same time (Baron-Cohen et al., 2001b).

The assessment itself consists of 36 photographs of actors' eye regions, with participants directed to select which one of 4 mental state terms best fits the picture. An example of one of these pictures that makes up the assessment's stimuli is included below in Figure 2. The test is scored on a scale of 0 – 36 with each correct answer scoring a single point. The average score of a neurotypical individual ranges between 26-28 points, with a standard deviation of between 3-4 points according to Baron-Cohen et al. (2001a). Participants have no time limit to answer the assessment's questions and can double check a list of mental state terms at any time in the original test, though some form of time limitation is often added to computerised versions of the test in use today (Baron-Cohen et al., 2015; Redondo, & Herrero-Fernández, 2018). The test itself is easily attainable with all the materials required to run the assessment posted for free on the Autism Research Centre's (ARC) website.

Figure 2.

Example RMET-R picture

Note. Picture taken from question 4 of the RMET-R. Available from https://www.autismresearchcentre.com/arc_tests

In regard to its high level of popularity the test itself is quick and easy to conduct and adapt into a wide variety of research designs. The assessment has a long history of use, and has been used to research a wide array of cognitive processes, disorders, and traits theorised to have some relation to ToM such as; verbal IQ (Peterson, & Miller, 2012), religiosity (Vonk, & Pitzen, 2017), ageing (Pardini, & Nichelli, 2009), and personality traits (Vonk et al., 2015). The quick and simple to conduct nature of the assessment lends itself particularly well to the physical demands of neuroimaging studies and research with individuals who have neurological issues; as such the RMET-R remains one of the most popular assessments for this area of research (Dal Monte et al., 2014; Thye et al., 2018; Hamilton et al., 2017). The assessment is also noted to be less influenced by memory or executive impairments compared to other ToM instruments (Lagravinese et al., 2017), which encourages its use with individuals with such impairments. Adding to its popularity is the range of languages the RMET-R has been translated into, along with the comparative ease in translating it compared to its peers. Over two dozen translated versions of the test are immediately available on the ARC website and in the research literature (Sanvicente-Vieira et al., 2014; Adams Jr et al., 2014).

There are some major problems with the RMET-R which have become increasingly criticised over time. The test has been found to be correlated with a range of other cognitive abilities (Brent et al., 2004; Peterson, & Miller, 2012), and its reported reliability varies wildly between different studies(Chen et al., 2017; Baron-Cohen et al., 2015). The psychometric properties of the RMET-R have also been poorly investigated, with only a limited amount of information reported by the study's original authors and in follow up studies which raise suspicions about its true psychometric properties(Peñuelas-Calvo et al., 2019). In addition, the test uses static, third party stimuli with explicitly cued answers, which make its questions much easier than real-world social

situations (Canty, 2016; Purcell et al., 2013). An Item Response Theory (IRT) analysis of the RMET-R by Black (2019) notes that the assessment is an easy test that has little ability to successfully discriminate higher ability individuals in neurotypical samples despite its common use for this task.

The test has also been found to have a slight cultural advantage for western participants in a couple of cross-cultural studies (Dodell-Feder et al., 2020; Söderstrand, & Almkvist, 2012). This advantage has been attributed to the western cultural background of the photographs of the actors' eyes in the test. The RMET-R has also been found to affected by participants age, gender and level of education, with these three variables found to have a significant effect on RMET-R scores (Baron-Cohen et al., 2001a; Fossati et al., 2017; Kirkland et al., 2013). The assessment has also come under criticism for only assessing affective aspects of ToM (Domes et al., 2007; Vonk, & Pitzen, 2017), despite its common use as an overall measure of ToM ability and the claims of its original creators (Baron-Cohen et al., 2001a). Another noticeable issue with the RMET-R that is of particular interest to this study is that no major research has ever been performed into how the RMET-R is actually performed. This issue has been almost universally glossed over, and although of less immediate concern compared to some of the other issues with the assessment is still a noticeable gap of information. This information potentially allowing for assessment of differences in how the test is conducted between different groups of individuals and if alterations to the measurement effect how individuals perform it. A wide variety of research has already been conducted using the RMET-R to assess ToM differences between groups of individuals (Baron-Cohen et al., 2015; Dal Monte et al., 2014) and a number of manipulated versions of the assessment have been created (Redondo & Herrero-Fernández, 2018; Adams Jr et al., 2010) which could have benefited from this information.

Despite this, the test has been noted to excel as an affective measure of ToM (Mintah & Parlow, 2018), with a score of battery assessments and research studies choosing to perform the RMET-R alongside another cognitive ToM assessment in order to conduct more thorough assessments of ToM abilities (Balaban, et al., 2016b; Martory et al., 2015). However, the RMET-R is scarcely alone in regard to these issues, with many ToM assessments noted to have a range of problems, as will be discussed below.

Advanced-ToM assessment problems.

There are a number of significant issues that are commonly found within a range of ToM assessments; with criticisms of psychometric issues, a lack of standardisation, heavy co-activation with other neurological processes, and issues with poor real-world validity found throughout the research literature (Canty, 2016). Regarding the psychometric criticisms, the psychometric properties of many ToM assessments remain either unreported or poorly explored (Bosco et al., 2016; Canty, 2016). This lack of data raises doubts about their true psychometric properties, and the validity of a vast number of studies conclusions that have used these assessment's. There is also a notable lack of standardisation, with a variety of; alternative scoring systems, administrator protocols, and abbreviated versions of a number of the most commonly used assessments in existence (Canty, 2016). For example, the short stories test and the Frith-Happé animations both have multiple scoring protocols, with the exact variation in use often poorly reported by the study's authors (Short stories: Happé, 1994; White et al., 2009; O'Hare et al., 2009 Frith-Happé: Abell et al., 2000; White, Coniston, Rogers, & Frith, 2011). It is not unheard of for researchers to only partially run these assessments either. Some researchers randomly select a variety of questions from assessments such as the short stories and Frith-Happé animations to present to participants (Short stories: Rakoczy et al., 2018 ; Im-Bolter et al., 2016 Frith-Happé: Abell et al., 2000; Wang et al., 2016).

Many ToM tests have also been criticised for having high correlations with other psychological abilities. This is problematic as differences in these alternative abilities have been found to influence performance in a variety of ToM tests. Which raises doubts about the supposed validity of these assessments (Brüne, 2005; Canty, 2016). Several commonly used tests have been found to be significantly related to executive functioning (Ahmed, & Miller, 2011; Lecce et al., 2017), memory (Crane et al., 2013), general cognitive abilities (Brent et al., 2004), and verbal IQ and other language abilities (Im-Bolteret al., 2016). Language abilities in particular have been noted as especially difficult to differentiate from ToM abilities, as both abilities are very closely related (Balaban et al., 2016a).

Although the majority of assessments are reported to be able to successfully differentiate clinical samples of individuals with low ToM abilities from samples of neurotypical individuals many assessments are criticised as insensitive to more subtle differences. Significant differences between test scores and real-life social skills have been reported by a variety of researchers (Dodell-Feder et al., 2013; Scheeren, de Rosnay et al., 2013). Which has led to the suggestion that many of many of the available tasks inaccurately represent their subjects real-life ToM abilities (Canty et al., 2017; Westerhof-Evers et al., 2014). It has been argued that this is due to poor real-world validity in such assessments, with even the most recently developed assessments relying on observation based, third party stimuli with explicit answer cues (Canty, 2016). The use of these methods simplifies the tasks to the point that they are significantly easier than real-world situations, which are typically more ambiguous and involve rapidly changing contextual cues from multiple sources (Klin, 2000). Bzdok et al. (2012) also argues that such tasks are far less mentally arousing and motivating for participants. Which results in participants using a different variety of underlying neural processes and lower engagement compared to real-life social situations.

The common use of older ‘classical’ assessments exuberates these issues, as they generally have more difficulties with these problems. These assessments are accused of having problematically low levels of sensitivity, with next to no ability to assess for differences between higher functioning individuals (Livingston et al., 2019a). In part to counter this there is an ongoing trend towards increased technological involvement in ToM assessments to generate more valid stimuli, with the Virtual Assessment of Mentalising Ability (VAMA) by Canty et al. (2017), which represents the current pinnacle in this developmental trend. The introduction of such technological methods allows for the generation of more engaging stimuli and increases the real-world validity of the assessments, but also introduces other problems. Typically increasing the complexity, material demands, and the time required to conduct the assessment, while still leaving them easier than real-world social situations. These ‘classical’ assessments, despite their known issues, offer quicker run times and ease of adaptation, as previously noted, which encourages their continued use. With a slow embrace of the improved and more rigorously tested recently assessments by the wider research community (Livingston et al., 2019a; Green et al., 2005). However, over the last few years there has been one new major issue that has drawn attention, which the available assessments are particularly ill suited to deal with: compensation.

Compensation and ToM

Compensation is a well-known phenomenon in cognitive and neurocognitive research, however, the issue of compensation in ToM has only begun to be explored in any real fashion over the last couple of years, with only a limited scope of research performed (Hull et al., 2019; Livingston et al., 2019). However, over the last few years it has become increasingly apparent that compensation is a major issue for the supposed validity of many ToM assessments. The term ‘compensation’ itself lacks a universal definition with multiple overlapping terminologies existing across a variety of fields (Livingston & Happé, 2017;

Ullman & Pullman, 2015). The concept of compensation in the context of this thesis follows the definition originally put forth by Livingston and Happé (2017) for referring to compensation in neurodevelopmental disorders. Which defines compensation as: “the processes contributing to improved behavioural presentation of a neurodevelopmental disorder, despite persisting core deficit(s)at cognitive and/or neurobiological levels” (Livingston & Happé, 2017, p. 731). Compensation in this case refers to any behavioural, cognitive, or neural process that contributes to improved behavioural presentation, which in the case of ToM is improved mentalising and social abilities.

Benefits/Drawbacks of compensation

Compensation allows individuals with deficits to superficially improve their abilities and camouflage their underlying deficits, allowing them to disguise persisting underlying problems by functioning in a more neurotypical manner (Livingston & Happé, 2017). Just as an individual’s level of impairment can differ in severity an individual’s ability to compensate can also vary depending on their cognitive facilities and prior experiences. An individual’s behavioural presentation is determined by both the severity of their deficits and their compensatory abilities, which can lead to a broad range of behavioural improvement. The exact form of an individual’s compensatory abilities also varies from individual to individual and can involve multiple compensatory processes and techniques that run in tandem.

Due to this, individuals differ in their ability to compensate. Some forms of compensation are easier to perform, and certain difficulties are easier to compensate for than others. Livingston & Happé (2017) divide compensation into two levels of complexity; ‘shallow’ and ‘deep’. Shallow compensation involves more basic strategies and behaviours that allow individuals to superficially camouflage their deficits such as imitating others’

thoughts and behaviours. Deep compensation involves the use of alternative cognitive mechanisms that allow for near neurotypical levels of function.

Compensation does not allow for the full remediation of behavioural and cognitive difficulties. Instead it offers improved ToM abilities, that are slower, more fragile, less flexible, and more resource intensive than the ToM abilities in neurotypical individuals (Livingston & Happé, 2017). The exact degree of behavioural improvement depends on the level of compensation that takes place, with poorer shallow compensation suggested to offer significantly poorer improvements and frailer abilities. Shallow compensatory strategies are also suggested to be inflexible, as they react poorly to novel or unexpected situations and ambiguous social cues. These strategies rapidly break down under stressful situations or when the compensators are mentally fatigued or anxious. They are also less likely to involve underlying socio-cognitive compensatory processes, which makes shallow compensation more likely to be picked up in assessments. Deep compensators' abilities on the other hand are posited to be much more flexible and resistant to outside stressors, with some deep compensators theoretically able to compensate to the point that their abilities are externally indistinguishable to neurotypical individuals (Livingston & Happé, 2017).

Regardless of its exact nature compensation is very mentally taxing. Both conscious and unconscious forms of compensation require additional resources to run, which can rapidly exhaust an individual's finite supply. The complicated nature of social situations makes this issue worse for compensating individuals as they typically require individuals to divide their finite cognitive resources between multiple consecutive tasks. The already complicated nature of social communication combined with the continuous drain of cognitive resources for compensatory purposes makes social activities especially draining for such individuals (Livingston & Happé, 2017). Multiple studies have found that higher functioning individuals with ASD that partake in compensatory behaviours consistently report that using

them in social interaction is highly draining and commonly leaves them mentally tired (Hull et al., 2017b; Cage et al., 2018). A number of other negative aspects of compensation have also been suggested over the years. The draining nature of compensation, for instance, has been suggested by several researchers to partially account for the high level of comorbid mental health difficulties, such as depression and anxiety suffered by such individuals by a couple of researchers (Hull et al., 2019; Livingston & Happé, 2017). The camouflaging nature of compensation has also been suggested to further compound these issues by hiding them, which results in reduced access to clinical or mental services and other useful other forms of support (Cage et al., 2018).

Recent research findings suggest that previous research may have over-emphasised negative aspects of compensation, however. Livingston et al. (2019c) finding that, although mentally taxing, the majority of compensating individuals report that the use of compensation is vital for them to have fulfilling life experiences. These individuals placing a high value on their compensatory strategies and reporting that they are very successful. The study finding that individuals that balanced using compensatory strategies alongside playing to their personal strengths had the best outcomes.

Forms of ToM compensation

As previously noted, research into the role of compensation in ToM is still in its infancy, as only a limited amount of research has been performed. Despite this, a variety of specific behavioural, cognitive, and neurocognitive compensatory strategies and processes have been identified in the research literature predominately through studies focusing on individuals with ASD (Livingston et al., 2019b; Hull et al., 2019). A sizable number of individuals with ASD for instance are commonly found to consciously use a number of behavioural strategies to camouflage their lower social abilities. Examples include the

conscious imitation or mimicry of neurotypical individuals' social mannerisms regardless of their degree of understanding (Lai et al., 2011), purposeful suppression of self-identified abnormal behaviours, and the planned avoidance of complicated social events (Livingston et al., 2019b). These camouflaging behaviours have been found to be more commonly used by higher functioning individuals (Hull et al., 2019), and autistic females compared to males (Hull et al., 2017a; Lai et al., et al., 2015). Forms of such behavioural compensation have also been reported in the social circles and job preferences of high-functioning ASD individuals in adulthood, with such individuals tending to gravitate towards social circles and careers with lower social and stronger analytical demands (Johnson et al., 2015; Baron-Cohen et al., 2000).

The use of a number of alternative cognitive faculties to boost up or replace one's ToM abilities has also been found in the research literature. An individual's executive functioning skills, memory abilities, general, and verbal intelligence are all suggested to be able to help bootstrap or supplant an individual's deficit ToM abilities (Livingston & Happé, 2017). Evidence of these abilities found to be stronger in high functioning and late diagnosed individuals compared to their lower functioning peers in ASD research supports this position (Durrleman, & Franck, 2015; Lehnhardt et al., 2016; Scheeren et al., 2013). These individuals theoretically use such abilities to compensate for their lower ToM skills in a number of ways. Their abilities for instance have been theorised to help them recognise and develop plans to manage their deficiencies. This enables them to more easily use analytical, avoidant or constraining behavioural strategies, as mentioned above. Hull et al. (2019) for example suggests that such individuals can use these abilities to develop appropriate topics for conversation ahead of time. Strong executive functioning and memory skills have also been proposed to help individuals with deficits communicate with others by allowing them to

recall previously learned social rules instead of relying on intuitive ToM abilities (Ullman & Pullman, 2015).

Several neuroimaging studies have also noted patterns of atypical neural activation amongst high functioning individuals with ToM deficits, which has been suggested to be indicative of neurological compensation. These individuals' brains are theorised to be utilising atypical neural routes to bypass deficiencies in the normal ToM network (Livingston and Happé, 2017). The exact format of alternative activation has been found to differ between studies. Livingston and Happé (2017) theorising that neurological compensation for ToM deficiencies is idiosyncratic, differing between individuals. Some studies for instance report hyperactivation (White et al., 2014) while others report hypoactivation (O'Nions et al., 2014; Kana et al., 2015) of core ToM network regions in individuals with deficits, which respectively indicates increased and decreased mental effort in such regions. Other studies report the use of alternative neural activation in individuals with ToM deficits when compared to neurotypical individuals (Kaiser et al., 2010a; Kaiser et al., 2010b). This suggests the use of alternative neural pathways by such individuals' brains to bypass impaired pathways in the normal ToM network. Their brains rewiring themselves to compensate for deficiencies in these brain regions in a similar fashion to how individuals' brains can rewire themselves following damage (Kantak et al., 2012).

Regardless of the exact form it takes, compensation has become increasingly recognised by researchers over the past few years as fairly common amongst individuals with deficits. The strong value of social skills in society and the large number of benefits it provides strongly encourages the use of compensation amongst able individuals (Hull et al., 2019; Preckel et al., 2018). All individuals with ToM deficits who have the potential to do so theoretically using some form of compensatory strategies. Recent research suggests that compensation explains several unanswered phenomena seen amongst individuals with ASD.

This includes ‘good-outcome’ individuals who transition away from diagnostic criteria over childhood as their compensatory abilities increase, and late diagnosed individuals who suddenly seem to run into trouble in late adolescence or adulthood whose compensatory strategies can no longer disguise their underlying deficits (Livingston & Happé, 2017). In regard to the large male-female diagnosis difference in autism, recent research has hypothesised this may be because females are better compensating than males. This hypothesis has seen a fair amount of recent interest and has inspired a number of new avenues of research (Hull et al., 2019; Lai et al., 2019).

ToM compensation and ToM assessments

It has been hypothesised that compensating individuals are also able to pass ToM assessments with ease. Theoretically deep compensating individuals are able to pass the majority of available ToM assessments with little difficulty (Livingston & Happé, 2017). And even shallow compensators are able to pass some of the simpler and more commonly used ‘classical’ assessments. The use of compensatory strategies and processes allowing such individuals to successfully pass ToM assessments while still they still possess significant ToM deficits and real-world social difficulties, which invalidates the assessments (Livingston et al., 2019b). This idea is supported by recent research findings suggest that these compensatory strategies are commonly used by a considerable number of individuals with ASD, which are a common focus and the most commonly used clinical group in ToM research (Hull et al., 2019).

The comparable ease and highly correlated manner of ToM assessments with other cognitive processes theoretically make them particularly vulnerable to compensation. For instance, compensating individuals who can successfully camouflage their underlying deficits are likely to be able to easily pass assessments that are simpler than real-world social

situations. High correlations can be found between several assessments and a variety of cognitive abilities. Which suggests that compensation through these abilities can be an easy alternative way to successfully complete such assessments. Multiple assessments have been hypothesised in the research literature to be able to be passed in this manner, including; the RMET-R, strange stories, and faux pas test (Ahmed & Miller, 2011; Ahmed & Miller, 2013; Scheeren et al., 2013).

This hypothesis has seen increased interest over the last couple of years, with several researchers having theorised that compensation through a number of different means can be used to pass a broad range of ToM assessments (Livingston & Happé, 2017; Hull et al., 2019). However, research on this topic is still in its infancy. With no empirical assessment yet conducted into if/or how individuals with ToM deficits can use compensation to pass ToM assessments (Livingston et al., 2019b; Beaumont & Sofronoff, 2008).

Despite this lack of research, compensation is still considered to be a major problem to the validity of existing ToM assessments by this thesis's authors. Compensatory techniques are commonly used by a considerable number of individuals of major populations of interest in ToM research (Hull et al., 2019). And the comparable ease and highly correlated manner of ToM assessments with other cognitive processes theoretically result in only the more severely impaired individuals able to be validly assessed (Hull et al., 2017b; Livingston et al., 2019b). Due to these issues the current ToM assessments are not considered to be sensitive enough to validly assess individuals with ToM deficits who use compensatory methods.

This raises doubts about the appropriateness of their use and the validity of a span of findings utilising such assessments. Additionally, they could potentially cause struggling individuals, with less obvious difficulties to be overlooked in research, causing them to miss

out on useful support or intervention (Livingston & Happé, 2017). Consequentially, this makes a way to more accurately assess compensating individuals in ToM assessments or assess for evidence of the use of compensation within such assessments, of great value.

The Current Study

One potential method that may help improve available and future assessments and may offer a way to assess for empirical evidence of compensation, may lie in the addition of increased answer and stimulus time constraints. These additional constraints theoretically deny compensatory individuals the extra time they require to use their compensatory techniques or processes and increases the difficulty and cognitive demands of the task to the point that their more fragile abilities faulter. Which would theoretically cause compensatory individuals scores to more accurately reflect their true ToM abilities and respond in a more consistent fashion, improving the validity and reliability of the assessment.

The addition of multiple of these constraints to a single assessment may also be highly beneficial. As this would allow an assessor to also assess differences between the time and visual stimulus constraints. This potentially enables the generation of a clearer picture of an individual's ToM abilities by varying the assessment's difficulty. And allow for the search of evidence of the use of compensation within assessments. Individuals that perform significantly worse than a neurotypical sample under shorter constraints but reasonably well under the assessments original conditions likely using compensatory processes to pass the assessment. A way to search for empirical evidence of compensatory processes use within ToM assessments is of considerable interest as this has not yet been performed (Livingston et al., 2019b). In the same manner these constraints may also allow for the assessment of the presence of compensation within individuals through the generation of norms between different time constraints, though this is beyond the current scope of this study.

Some support for the viability of this additional methodology can be found in studies of compensation in other neurological conditions such as dyslexia in which the addition of time constraints (Parrila et al., 2007) and additional stressors (Varnet et al., 2016) has been found to slow down reading abilities. There is also research which has found increased latency when responding to ToM test questions amongst individuals with high levels of autistic traits (Miu et al., 2012). Miu et al. (2012) specifically finding that individuals who scored one standard deviation (SD) below on the AQ took significantly longer but performed no differently than individuals who scored one SD above on the RMET-R. The practice of limiting the presentation of stimuli has also long been used in memory research to increase the difficulty and cognitive demands on a variety of tests (Baddeley, 2003; Bachmann & Francis, 2013). This form of additional constraint theoretically working the same way for ToM assessments.

The inclusion of these additional constraints also allows for the manipulation of the assessment's difficulty and increases the test's real-world or ecological validity. As previously noted, current ToM assessments are significantly easier than real-life social situations which typically require far faster responses and involve other cognitive processes that need to run in tandem with an individual's ToM abilities. Both cognitive and affective assessments of advanced ToM should theoretically benefit from these additional constraints, though this current study is only designed to assess an affective assessment. Hypothetically due to the differing nature of affective and cognitive ToM processes the time constraints for affective measures would need to be significantly smaller than cognitive ones, and vice versa for the visual limitations for cognitive assessments though this would require further research beyond the immediate scope of this study. As previously mentioned, there isn't considered to be much value in adding these constraints to simple ToM assessments as they too simple for older individuals that are more likely to be using compensatory processes. However, there

may be some value in manipulating the difficulty of such assessments to gather a more accurate picture of ToM development, though this is beyond the immediate focus of this project.

Despite the simplicity of these two additional constraints and their use in other cognitive assessments, to our knowledge the use of multiple answer and stimulus time constraints have not been used in ToM assessments before. Singular time restrictions are commonplace in the majority of newly developed assessments and are commonly added to a number of older 'classical' assessments today but the value of increasing the time restrictions to manipulate the test itself has not been widely explored. These constraints can easily be added to available assessments, and are quick and easy to perform, requiring few additional resources. This makes them highly viable for both research and clinical purposes and increases their probability of adoption in research and clinical conditions where time and resources limitations can easily exclude the use of more complicated assessments. Many of the more recently developed assessments are unfortunately limited in these manners. Their increased focus on higher tech designs and increased ecologically valid stimuli unfortunately cause them to take longer, require more resources and are more difficult to adapt to research demands which restricts their use (Livingston et al., 2019a). Due to this, these additional constraints are of great theoretical value for both existing and future assessments.

Outline of the current study

As such, the main focus of this current study is to assess the potential viability and validity of adding additional answer and stimulus time constraints to a ToM assessment. The end goal of which is to assess if these constraints can increase the assessment's psychometric abilities and allow one to assess for evidence of compensation used to pass the assessment. For this study the RMET-R was selected to be used due to its position as the most popular

ToM assessment: furthermore, it is an older, static, explicitly cued task that lacks any kind of time limit and has been found to be correlated to a couple of other cognitive abilities. This theoretically makes it easy to pass using a number of compensatory strategies and marks it as prime task for improvement.

Based on the results of a couple of prior pilot studies, we chose to run a between subjects' experiment. Participants were randomised between three purposefully designed timed variations of the RMET-R, where a failure to answer in time counted as an incorrect answer. These three variations make up the experiment's three experimental conditions and consist of; a 'long' variation, with a 20-second answer time limit; a 'short' variation, with a 5-second answer time limit; and an 'occluded' variation, where in addition to a 5-second answer time limit the tests pictures were occluded by a mask of static after half a second in a manner akin to stimulus limitation practices used in masking in working memory tasks (Ricker, & Sandry, 2018). The long variation was used in this study as a stand in for the original assessment, while the other two conditions were used to assess the effect of the two constraints. The short variation assessed the effect of further limit the available answer time, and the occluded variation assessed the effect of limiting the stimulus presentation.

As we lacked access to a clinical sample of individuals with poor ToM abilities, we chose to also assess two other related constructs to ToM to assess if the addition of these constraints could improve the concurrent validity of the RMET-R. With Empathy and mentalizing eventually selected. Empathy was chosen to be assessed using the Toronto Empathy Questionnaire (TEQ) developed by Spreng et al. (2009) and Mentalising by the Mentalization scale (MentS) developed by Dimitrijević et al. (2018). Both of these constructs found to be positively correlated with ToM in previous research, which means individuals with poorer scores in these measures are more likely to have poorer ToM abilities. A short version of the AQ; the AQ-10 was also investigated as a potential assessment, as measures of

autistic traits such as the AQ have been commonly used for this purpose in previous research. However, we chose not to use the AQ-10 in our main study, based on of the results of a pilot test which contrary to as reported in the research literature found no relation between AQ-10 and RMET-R scores.

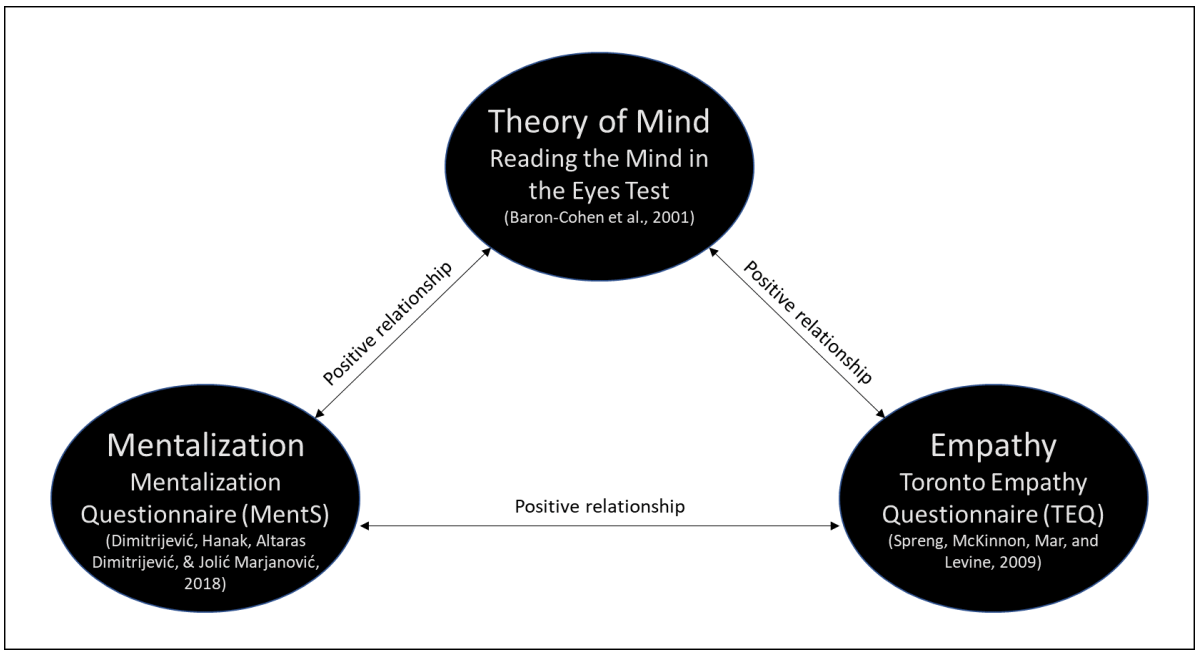
Though empathy and the TEQ itself has been found to be positively correlated with the RMET-R in previous research (Spreng et al., 2009) it should be noted that there is ongoing debate over empathy's true relation to theory of mind (Johansson Nolaker et al., 2018; Preckel et al., 2018). Empathy is considered by this study's authors to be correlated with at least affective ToM, as is assessed by the RMET-R. As empathy has been found in multiple studies to be closely neurologically related to affective ToM, as it involves an overlapping neural system (Volm et al., 2006). A couple of researchers even suggest co-activation of both processes when inferring others' immediate mental states (Lamm et al., 2011).

The mentalization scale on the other hand is a self-answer measure of mentalization, which is a very closely related concept to ToM. The only major difference between the two is that the conceptualisation of mentalization places a greater focus on the ability to self-reflect on mental states (Dimitrijević et al., 2018). The MentS even includes a subscale of other-related mentalising abilities which for all purposes is effectively a self-report ToM assessment. The MentS has also been found to be positively correlated with the Cambridge Behaviour Scale Abbreviated Empathy Quotient (EQ) in Dimitrijević et al. (2018). The EQ is another measure of empathy, which has been found be correlated with both the TEQ and RMET-R in multiple studies (Spreng et al.,2009; Trent et al., 2016). Due to this the assessment measure, or at least its other subscale should theoretically have a positive relationship with affective ToM assessments such as the RMET-R, though this has not yet

been assessed in the research literature. Figure 3 below demonstrating the theorised relationship of the three main assessments of our study.

Figure 3.

Theorised relationship of the three main constructs in this study



Aims of the study

Main Aims

The primary aim of this study is to assess if shortening the answer or stimulus presentation time in the RMET-R can increase its difficulty, validity and reliability. These additional constraints theoretically disrupt or remove the ability to use a range of compensatory processes used by individuals with poor ToM skills. This increasing the difficulty of the test and theoretically leading to an improvement in the test's construct validity and internal consistency.

To answer these questions this study uses the three manipulated variations of the RMET-R which make up the study's three experimental conditions and levels. With the study specifically interested in if the relationship between RMET-R scores and TEQ and MentS

scores differs between the three variations. The presence of a stronger positive relationship between RMET-R scores and MentS/TEQ scores in the short variation compared to the long supporting our theorised value of increased answer time constraints. And the presence of a stronger positive relationship between RMET-R scores and MentS/TEQ scores in the occluded condition compared to the short and long conditions supporting our theorised value of stimulus time constraints.

In addition, we are also interested in assessing (1) if the RMET-R scores differ between the three variations, to ensure that our manipulations effected the test as we expected it to. And if internal consistency as assessed by Cronbach's alpha differs between the three variations, with higher levels indicative of improved reliability.

Exploratory aims

This study is also interested in exploring two secondary exploratory aims: (1) conducting a simple slopes analysis at the+/- 1 SD level of the three RMET-R's variations relationships to the TEQ/MentS if a statistically significant interaction is found between them, and (2) assessing what methods people report they use to determine the RMET-R's answers. The former aim would help us as we are interested in investigating whether these constraints may also allow one to search for initial evidence of the use of compensation to pass assessments. If compensation can be used to pass such assessments individuals that likely have poor ToM abilities, as indicated by lower MentS and TEQ scores, should theoretically be able to pass the long variation without much difficulty. However, they should also struggle in the two shorter variations when they are rendered unable to use their compensatory abilities, allowing for individuals to search for this pattern of results. As stated earlier multiple assessments have been assumed to be able to be passed through the use of

compensation, though no empirical assessment has yet been conducted (Livingston et al., 2019).

Meanwhile the latter aim would assist in the investigation of if there are any relationships between the use of any specific or amount of methods used, and RMET-R, TEQ or MentS performance. This is of interest to the authors of the study as we are interested in assessing what strategies individuals use to complete the RMET-R. This aspect of the assessment has never been investigated before and may be of considerable value for assessing differences between groups of individuals. For example, compensating individuals should theoretically use different strategies to complete the RMET-R than neurotypical individuals. Compensating individuals possibly relying on less flexible strategies or increased guesswork while neurotypical individuals may be able to more easily swap between different strategies depending on the demands of the question. As such, we are interested in investigating how individuals complete the assessment and assessing if there is any relationship between the use of any specific or amount of methods and RMET-R, TEQ or MentS scores. It is also quite possible that our experimental manipulations may affect how individuals complete the assessment. The rise in difficulty and theoretical effect on compensatory processes caused by our additional constraints potentially restricting the use of specific strategies and limiting individuals' abilities to use multiple strategies. Due to this we are also interested in assessing if any of our experimental manipulations change how individuals perform the test.

Hypotheses

The hypotheses for the current study are divided into two sections. One focussed on the hypothesized relationship between the three RMET-R variations scores and TEQ scores. And the other on the relationship between the three RMET-R variations and MentS scores. However, these two sets of hypotheses are identical aside from the measures involved,

making it more practical to discuss them together to avoid repetition. As shown in the previous chapter both Empathy (TEQ) and Mentalizing ability (MentS) are both theorised to have a positive relationship with affective ToM. Which if found to be correct in our experimental sample will allow us to assess if our additional constraints improve the validity of the measure by improving the relationship between the RMET-R and the TEQ and MentS.

Two separate constructs that were both reported to be related to ToM were chosen to be assessed in this manner to help avoid drawing any erroneous conclusions about the validity of these additional constraints' ability to improve the construct validity of the RMET-R due to noise in the data. As well as proving us with a redundancy in case our assumption about one of the measures' positive relationship with the RMET-R was found to be incorrect. Due to this possibility we had to begin by making sure the TEQ and MentS correlates with the RMET-R in the manner we theorised they do. Although both concepts theoretically correlate as explained, the exact nature of the relationship in our experimental sample is still unknown. Therefore, we hypothesised that in line with previous research:

Hypothesis 1: RMET-R scores will be positively correlated with TEQ scores.

Hypothesis 2: RMET-R scores will be positively correlated with MentS scores.

If these initial hypotheses are found to be correct, we can then reliably assess how our additional time constraints influence the RMET-R. At least one of these measures is required to correlate in their hypothesised manner to be able to continue forward with our planned analyses.

In relation to the addition of increased time limitations to the RMET-R in line with our theory we predicted that if Hypothesis 1 and/or 2 was correct:

Hypothesis 1A: RMET-R scores will more strongly covary with TEQ scores when RMET-R trials are shorter (5s) than when the trials are longer (20s).

Hypothesis 2A: RMET-R scores will more strongly covary with MentS scores when RMET-R trials are shorter (5s) than when the trials are longer (20s).

This pattern of results would provide evidence of stronger concurrent validity amongst the short answer time condition compared to long answer time. Which would support our theorised value of adding increased answer time constraints to the assessment along with adding some preliminary support for the proposition that the use of multiple answer times would allow for the assessment of evidence of compensation.

Following our theorised value of adding additional stimulus visibility constraints to the RMET-R we also predicted that if Hypothesis 1 and/or 2 was correct:

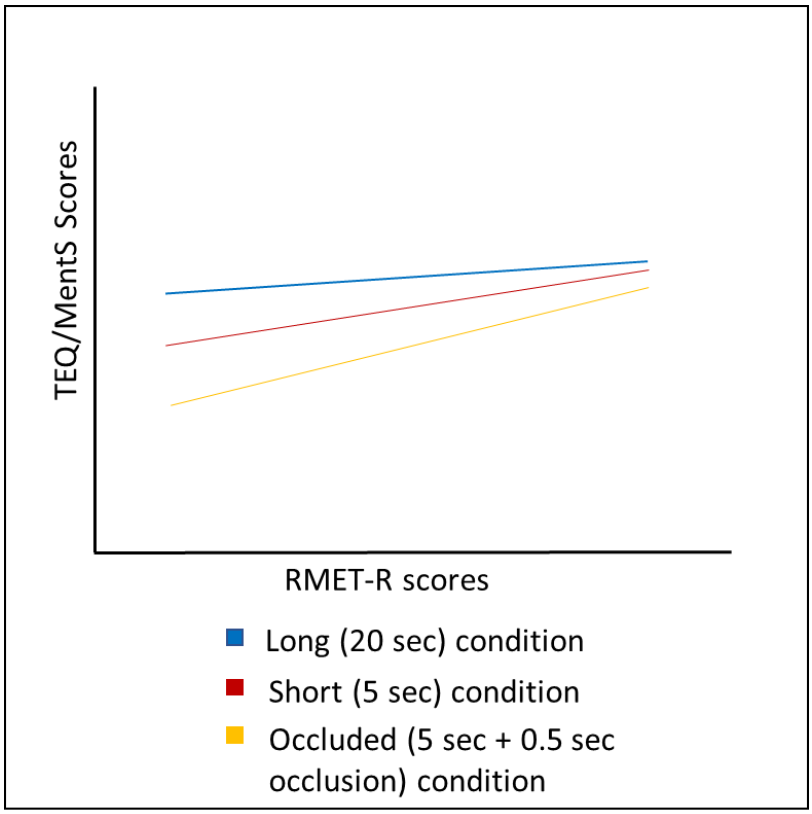
Hypothesis 1B: RMET-R scores will more strongly covary with TEQ scores when RMET-R trials have been occluded (seen for 0.5s, 5 s to respond) than when the trials are not occluded (5s and 20s trials).

Hypothesis 2B: RMET-R scores will more strongly covary with MentS scores when RMET-R trials have been occluded (seen for 0.5s, 5 s to respond) than when the trials are not occluded (5s and 20s trials)

This pattern of results would provide evidence of stronger concurrent validity in the occluded condition compared to the other two conditions. Which would support our theorised value of adding increased visual stimuli constraints and adding some more support for their use to assess for evidence of compensation. Figure 4 below illustrates the hypothesised relationship in hypothesis 1A, 1B, 2A and 2B between the Long, Short and Occluded RMET-R conditions and the TEQ/MentS.

Figure 4

Hypothesised relationships between the Long, Short and occluded conditions RMET-R scores conditions and TEQ/ MentS scores



Although we did not include it as a main hypothesis in the pre-registration, we are also interested in assessing if our experimental constraints increase the difficulty of the RMET-R by conducting a one-way ANOVA to examine if there is a significant difference in RMET-R scores between our three conditions prior to conducting our regression analyses. With significantly lower RMET-R scores hypothesised in the occluded condition compared to the short and long conditions and in the short condition than compared to the long.

We are also interested in analysing the internal consistency of each of the RMET-R conditions by calculating a Cronbach's alpha for each variation. A higher Cronbach's alpha is hypothesised in the occluded compared to the short and long conditions, and in the short compared to the long condition. Theoretically the more limited variations increase in

difficulty and interference with compensatory processes cause individuals to answer the test questions in a more consistent manner depending on their core ToM abilities.

Chapter 2

Method

This following chapter presents an overview and description of the current study's methodology. It includes details about the study's pre-registration, design, participant recruitment, dependent variables, measures and materials used, and the procedure of the experiment.

Pre-registration

Prior to commencement of data collection, a pre-registration outlining the methodology of the current study was created using the AsPredicted.org outline and presented on the Open Science Framework (OSF; https://osf.io). The pre-registration outlined the study's hypotheses, dependent variable, experimental conditions, planned analyses and sample size, outlier and exclusion criteria, along with some planned exploratory analyses. In addition, the pre-registration also contained a link to a Google Slides document demonstrating the full design of the experiment. The pre-registration was made public on the 3rd July 2019 and can be accessed at: https://osf.io/ykupf

Design

An experimental between-subjects design was used for this online study, with participants randomly assigned to one of the three experimental RMET-R conditions; 20 second time limit, 5 second time limit, or occlusion. Participants also received the TEQ and MentS in a random order to help avoid possible order effects. Participants randomised between these different conditions through the use of the (randomize = (function ()) function in java. The independent variables of the study were the three-time restriction variations of the RMET-R. The dependent variables of the main hypotheses for this study included the

RMET-R scores, TEQ scores (assessing empathy level), and MentS scores (assessing mentalising ability).

Sample size determination

Before commencing with this study, a linear multiple regression: fixed model, R2 increase a priori power analysis using G*Power version 3.1.9.2 (Faul et al.,2007) was conducted to determine the required sample size for this study. The analysis had an effect size of 0.07, and a power of .90 (α= .05), and a total and tested number of predictors of 2, indicating that a sample size of at least 184 participants was required. The effect size of 0.07 was selected based off the results of previous pilot tests which suggested that we wanted to be sensitive enough to detect a change in R2 of at least .06. As we theorised that adding increased answer and stimulus presentation times to the RMET-R would improve the validity and reliability of the assessment by stopping compensating individuals from using their compensatory strategies and lacked access to a clinical sample we selected to boost up our sample size further. As such, we aimed to have a sample size of at least 330 participants with at least 100 participants in each condition to increase our chances of including compensating individuals in our study.

Participants

As we lacked access to any kind of clinical sample we instead elected to use a large non-clinical online sample of individuals for our experiment. Participants were selected and recruited through the crowdsourcing platform website Prolific: https://www.prolific.co/. Which is occasionally referred to by its prior name of Prolific Academic and is named such in this study's pre-registration.

A convenience sample of 381 participants (179 M, 190 F, 11 gender diverse) with a mean age of 28.8 (SD=7.52) were recruited online from Prolific. The study was set up on

Prolific to only show up for individuals who reported they were between the ages of 18-45 and resided in the United Kingdom, United States, Australia, or New Zealand, and who had not participated in any previous related pilot studies by this study's authors. All participation in this study was voluntary with participants giving informed consent prior to the commencement of the study. The study had an estimated completion time of 20 minutes with a maximum allowed time of 35 minutes. Participants were randomly assigned to one of the three RMET-R time conditions (20 sec, 5 sec, occlusion). After the completion of the study participants received a monetary reward as compensation for their participation in the study. Participants receiving £ 1. 80 (approx. NZD$3.38 at time of study) per hour of time spent on the study, which was paid out in the form of credit on Prolific which can be cashed out via a PayPal account. Table 1 below displays the participant demographics for each of the study's three conditions, and the total experiment.

Table 1.

Gender and Mean Age of participants prior to exclusion checks in the long, short and occluded conditions of the experiment and the total experiment.

Demographics	**Long condition**	**Short condition**	**Occluded condition**	**Total experiment**
***Gender*:**				
Male: *n* (%)	56(45.2%)	66(52.8%)	58(43.9%)	179(47.1%)
Female: *n* (%)	65(52.4%)	57(45.6%)	68(51.5%)	190(50%)
Gender diverse: *n* (%)	3(2.4%)	2(1.6%)	6(4.5%)	11(2.9%)
Age: **Mean (SD)**	29.3(7.72)	27.7(7.24)	29.6(7.48)	28.8(7.52)

Note. N = number of participants, % = percentage of condition, SD = Standard deviation

As we aimed to have a sample size of 330 participants with at least 100 participants in each condition we initially setup the experiment for 330 individuals on Prolific, from which we received 323 datasets. After going through exclusion checks we were left with datasets from 293 individuals, with 94 participants in the long condition, 91 participants in the short condition, and 111 participants in the occluded condition. Due to this we reset the study on Prolific to run another 60 participants from which we received another 58 datasets, 3 further of which failed exclusion checks. This brought up the total number of participants assessed up to 381, with 30 having failed exclusion checks.

Measures

General study information

This experiment was conducted online via the use of Prolific, with all the test's measures and materials run through a single java program. In the program participants choose when to move forward on each page by clicking a button at the bottom of each screen in every section of the experiment apart from the experimental RMET-R task which is described in more detail below. Apart from the RMET-R task participants were also required to fulfil a two-step authorisation process to move forward if they had not provided an answer for every question. A textbox appeared for such individuals upon clicking the button to go to the next page, informing them they had not answered every question and asking if they were certain they wanted to continue. A selection of yes allowed them to continue onwards and a selection of no returned them back to their current screen. This was done to help participants avoid accidently missing any questions. This enabled us to lower the chance of receiving data sets with missing data, while not restricting their ability to choose to not answer any specific questions.

Prolific entry information

As participants were drawn from Prolific for this study, this study originally appeared on the website under the title of 'Initial Impressions from Eyes (3)' for those who met the entry criteria as described in the participants section. Before participants could commit to taking part in the experiment, they were provided a basic description of the study on Prolific. This basic description informed them that in this experiment they would be directed to complete a series of tasks including a demographic questionnaire and a task in which they would answer a series of questions based on a series of images of the eye region of a face. Participants were informed this experiment was designed to be conducted on a desktop computer or laptop and had an estimated completion time of 20 minutes with a maximum allowed time of 35 minutes. Those that elected to take part in the study were provided a link which brought them to a webpage housed in Massey University's psychology lab server which contained the main experiment. Upon completion of the experiment the participants were provided with a link that once clicked brought them back to Prolific and assigned them their agreed upon credit for completing the study.

Reading the Mind in the Eyes Test Revised (RMET-R)

The Reading the Mind in the Eyes Test Revised (RMET-R) is a theory of mind assessment developed by Baron-Cohen et al, (2001a) that was designed to be sensitive enough to assess deficits in social cognition amongst adult populations. The assessment consists of 36 photographs of the eye regions of different actors, with the participants asked to choose which one of four mental states best describes what the eyes are showing. The test is scored on a scale of 0 – 36 with each correct answer scoring a single point, higher scores indicating higher ToM abilities. Average scores from neurotypical sample range between 26-28 with a SD of between 3-4 according to Baron-Cohen et al. (2001a). The assessment also includes a single practice question for participants to complete before they commence the assessment, and a list of the mental state terms used in the assessment along with an

accompanying definition and example. The measure is available for free from the ARC website at: https://www.autismresearchcentre.com/arc_tests. However, as this thesis revolves around assessing if we can improve the psychometric properties of this assessment our three variations have a number of differences to the base assessment.

Difficulty Manipulation.

As we are using three manipulated variations of the RMET-R that we purposefully designed to be to be used online, our variations of the RMET-R all differ from the standard assessment. The most notable differences are the additional answer and stimulus time constraints alternatively added to our three variations. The Long variation including an additional 20-second answer time limit, the Short a 5-second time limit, and the Occluded a 5-second time limit along with a half a second stimuli presentation time limit - where the pictures that make up the RMET-Rs' stimuli were occluded behind a mask of white noise after half a second. In each of these variations a failure to answer in time counting as an incorrect answer.

The Long variation was used as a stand in for the original assessment, despite the original assessments' lack of a time limit. A time limit was sought for this 'Long' condition to help streamline the experiment and lower the amount of time participants may potentially spend performing it. A 20 second limit was considered a valid stand for the original test based off the results of a prior pilot study which found no differences in scores between a 20-second and a 2-minute time limit variation of the RMET-R. Further supporting this decision, a number of recent online studies using the RMET-R, including some of this measure's original creators (Baron-Cohen et al., 2015), have also added a 20 second time limits to the assessment.

The Short variation was used to assess the effect of limiting the available answer time, to assess the theoretical value of this additional constraint. Five seconds was selected as the most apt time limit, based on the results of a couple of prior pilot studies which suggested that low empathy individuals began to show significant difficulties compared to higher functioning individuals at this level.

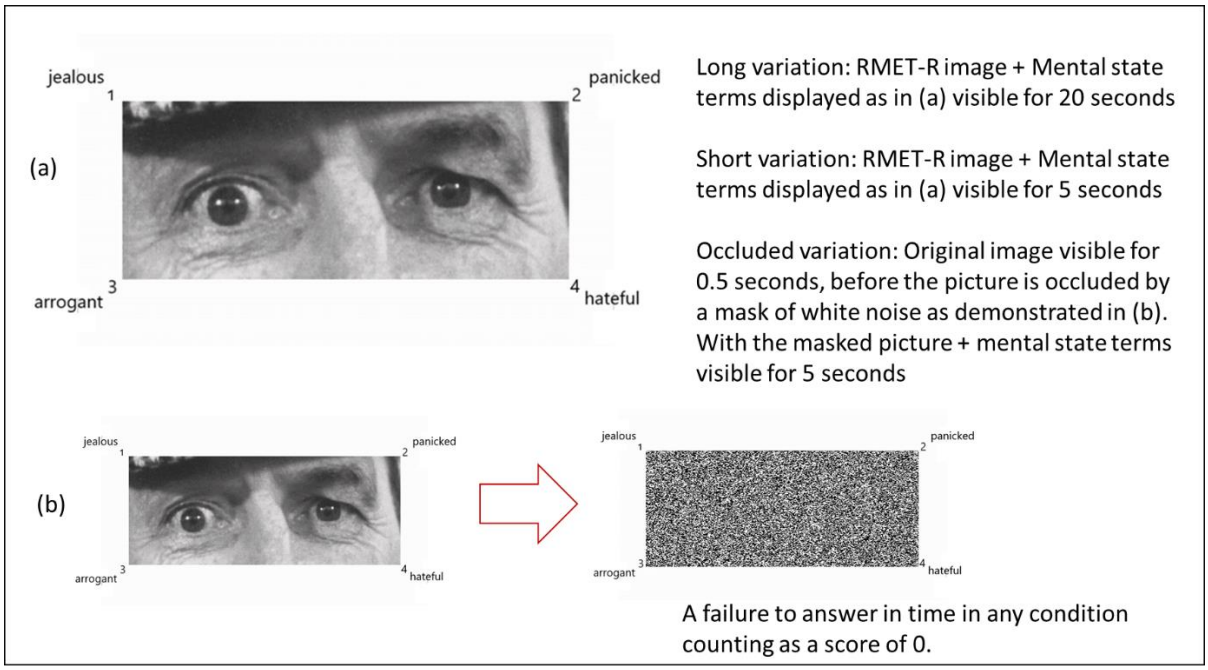
Finally, the Occluded variation was used to assess the effect of limiting the stimulus presentation time. In this condition in addition to a five second answer time limit the RMET-R's pictures were occluded behind a mask of static after half a second, while the pictures' four possible mental state terms remained visible. This form of stimulus limitation was chosen due to its common use and proven value in increasing the cognitive difficulty of memory and attention measurements (Bachmann & Francis, 2013). A short pilot test conducted prior to the commencement of this study also found that this form of stimuli constraint increased the difficulty of the RMET-R in the theorised manner but could still be accurately answered by neurotypical individuals.

Online variations.

To disguise the test in our study we the name "Initial Eye Impressions Task" to refer to the test throughout the experiment. To allow participants to quickly scan and distinguish between the available answers for each of the RMET-R's questions we chose to present the four possible mental state terms alongside their corresponding picture. Each of the terms were placed in a separate corner of the screen and paired with a number key ranging from 1-4, with participants directed to select the mental term that best fit what the individual in the picture was feeling using the number keys at the top of the keyboard. Figure 5 below illustrates how the RMET-R was presented in each of the three variations used in this study.

Figure 5.

Illustration of the presentation of RMET-R images and mental state terms in the three variations used in this study.



A full copy of the instructions, pictures, and answers used in this study is included in Appendix A. In contrast to the original RMET-R where participants were presented with a list of mental state terms that included a definition and example of each that they could to refer back to at any time, our computerised task presents this information to participants before they begin answering the RMET-R questions. Participants are instead asked to scan a list of mental terms and click on any that they are unfamiliar with. Once clicked the definition and example of the term taken from the original list is displayed. A full copy of the list of mental state terms and their accompanying definition and example is included in Appendix B. Figure 6 below illustrates how the mental state terms were presented in the study.

Figure 6.

Illustration of how the RMET-R mental state terms list, definitions, and example was presented in the study.

Manipulation check questions

To examine if the experimental manipulations of our study increased the difficulty and mental effort required to complete the assessment in the manner we expected, we also included three self-report rating scales as a manipulation check. These self-report ratings were situated directly after the RMET-R in all three conditions and consisted of three 10-point Likert scales. This short Manipulation check was paired with the second attention check question in the experiment. The questions and related scales were as such:

- ‘How difficult did participants find the task?’, along with a scale ranging from extremely easy (1) to extremely hard (10).
- ‘How much mental effort was required for the task?’, along with a scale ranging from extremely little effort (1) to extremely large effort (10).
- ‘How well did they think they did on the task?’, along with a scale ranging from extremely well (1) to extremely poorly (10).

The full text and design of this set of questions is included in Appendix C.

Toronto Empathy Questionnaire (TEQ)

The TEQ is a brief self-report measure of empathy created by Spreng et al. (2009). The measure was specifically developed to provide a broad, single construct measurement of empathy. The TEQ is composed of 16 statements that the assessments participants are instructed to carefully read and honestly rate how frequently they feel or act in the described manner using a five-point scale. The measure has a total score of 64 with a mean score of 45 found in the original validation of the questionnaire, with higher scores indicative of stronger empathy. In this experiment the TEQ is presented in an unmodified format, with its scores used as a moderator variable. All the materials to run the TEQ can be found in Spreng et al. (2009) with a full copy of the measure included in Appendix D.

The Mentalization Scale (MentS)

The MentS is a recently developed self-report measure of Mentalising ability created by Dimitrijević et al., (2018). The assessment is designed to assess an individual's ability to interpret other individual's mental states, their own mental states, and how often they reflect upon them. The assessment is made up of 28 questions each of which consists of a statement that participants are directed to carefully read and use a 5-point scale ranging from completely incorrect (1) to completely correct (5) to report how correctly each statement describes them.

The measure has a total score out of 140 and is made up of three subscales: a Self-related Mentalization (MentS-S), an Other-related Mentalization (MentS-O), and a Motivation to Mentalize (MentS-M) sections. The MentS-S consists of 8 questions (Q: 8, 11, 14, 18, 19, 21, 22, 27), the MentS-O 10 questions (Q: 2, 3, 5, 6, 10, 12, 20, 23, 25, 29) , and the MentS-M 10 questions (Q: 1, 4, 7, 9, 13, 15, 16, 17, 24, 28). Higher scores on the

assessment indicating higher mentalizing abilities. Everything required to use the MentS is available in Dimitrijević et al. (2018), and a full copy the measure is included in Appendix E.

In this experiment the MentS will be used as a moderator variable and as previously noted is theorised by this study's authors to positively correlate with RMET-R scores. In this experiment a single attention check question was also included between the 25th and 26th questions of the MentS which is discussed in further detail in the attention check section.

Marlowe-Crowne Social Desirability Scale Short-Form (MCSDS-SF)

To help improve the validity of our self-answer responses in our study and to help limit self-desirability-bias we included a newly developed short form of the Marlowe-Crowne Social Desirability Scale (MCSDS), originally developed by Crowne & Marlowe (1960). The MCSDS-SF is a new version of this assessment, specifically re-developed by Vésteinsdóttir et al., (2017) for internet surveys, making it ideal for our online experiment. The MCSDS-SF is composed of 10 statements where participants are asked to note whether each statement is true or false as it pertains to them personally. Higher scores are indicative of stronger tendencies to answer in a socially desirable manner. A copy of the measure's materials included in Appendix F.

The MCSDS was developed to detect individuals that are likely to 'fake good' in self-report measures. Such individuals purposefully answer in a socially desirable manner, exaggerating their positive traits so as to be viewed more favourably. This is a major problem with self-answer assessments, as this behaviour invalidates the individual's data. This has led to the development of a number of measures designed to indicate individuals that are more likely to engage in such behaviour (Lambert et al., 2016). The decision to include the MCSDS-SF in this experiment was made due to the presence of some unusual results in previous pilot tests, and the fact that empathy and ToM skills are typically highly valued.

Despite its advanced age the MCSDS has been found by Lambert et al. (2016) to still outperform more modern measures of socially desirable responding.

Self-report questionnaire of methods used to assess RMET-R stimuli

As previously mentioned in the course of developing this experiment, we also became interested in assessing what methods were used to complete the RMET-R and whether there was be any connection between them and RMET-R/ToM ability. As next to nothing in the research literature has investigated how the test is actually completed by participants we elected to design a small self-answer questionnaire to investigate what methods were used to complete the RMET-R, in what manner they are used, and if there are any relationships between them and the other assessments in this study.

Following standard data analysis techniques used in qualitative research reported in Kawulich (2004) and Thorne (2000) we initially came up with five main strategies we believed participants used to complete the test. These strategies were based off of the results of some preliminary data from an earlier pilot test in which we directly asked individuals how they believed they had completed the RMET-R and prior research on how affective ToM abilities and the RMET-R function. These five main strategies including:

1). The use of Intuition by participants, with participants who admitted they were unsure how they performed the test also added into this category.

2). Imagining the rest of the face, with participants relying on their visual imagery skills.

3). Comparing the pictures of the actor's eyes to their own previous experiences.

4). Rules of thumb, with participants basing their decisions on the direction of the actor's eyes, level of squint, etc.

5). Relying on their knowledge about the mental state terms to rule out obviously wrong answers and make a best guess, with participant relying on their verbal intelligence.

However, in our main study we elected to directly ask participants after completing the RMET-R what the main method or technique they believed they had used to complete the assessment was and provide a space for participants to write down their answer. This open-ended format was used to help avoid influencing participants and issues related to experimenter bias. This design also allowed us to us to investigate for the presence of other alternative methods that we had not yet come across in previous testing, as although the data from our pilot studies seemed to support our five strategies format the sample size only consisted of 59 individuals.

Following this, as we were also interested in assessing if individuals commonly used multiple strategies we chose to assess if participants also used any secondary techniques. To simplify the analysis of this section and help participants to recall if they used any secondary strategies, we chose to present a list of the five major strategies that we theorised were in use. Participants were then asked to indicate whether they used any of the five strategies in addition to their previously noted main method. Afterwards to help determine how participants used their strategies participants were then asked to indicate whether they:

- only used a single main strategy
- switched between individual strategies depending on difficulty
- commonly used a combination of strategies to reach an answer
- or used a combination of strategies for difficult judgements.

The full text and design of this questionnaire is included in Appendix G.

Attention checks

To help catch inattentive subjects and improve the power of this study two different kinds of attention checks were added to this experiment. The first one was a modified honesty check question designed to help catch inattentive and malicious responders which states: "I once owned a three headed dog". This question was hidden between the 25th and 26th questions of the MentS in this experiment and paired with a five-point scale akin to the scale used in the MentS to help disguise it. The scale ranged from completely incorrect (1) to completely correct (5), with any score other than a 1 counted as failing the attention check.

The second attention check included in the experiment was the first question of the Directed Questionnaires Scale (DQS). This single question from the scale was chosen to be included as its creators, Maniaci and Rogge's (2014), found that including even a single one of the DQS's questions can reliably improve a study's power. With the first question found to be the most effective of the scale. The question itself states: "I read instructions carefully. To show that you are reading these instructions, please leave this question blank." In the experiment this 2nd attention check was placed directly after the three manipulation check questions and paired with a similar 10-point scale ranging from extremely true (1) to extremely false (10) to disguise it. The provision of any answer by participants to this question counted as failing the attention check. In this experiment a failure of both attention checks was counted as an exclusionary criterion, with participant's that failed both attention checks datasets excluded from all data analyses.

Demographic questionnaire

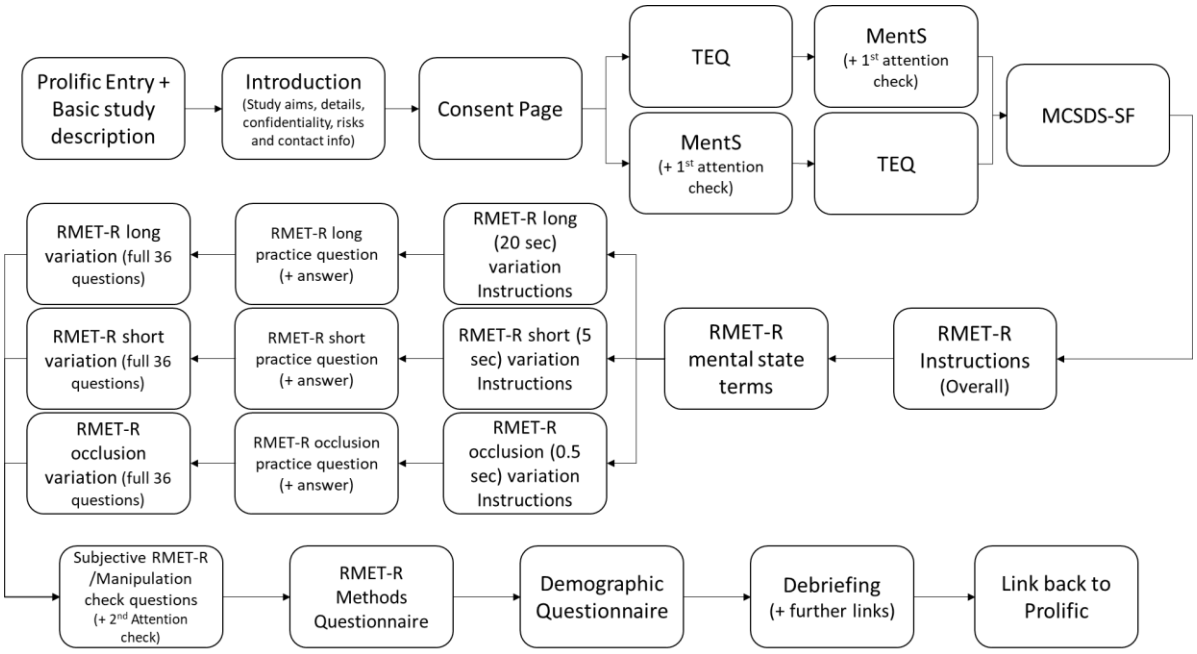
The experiment also included a demographic questionnaire at the end of the experiment created by the experiment's authors. The questionnaire assessed participant's; age, gender, language, education level, physical and mental difficulties, and familiarisation with the RMET-R. The full text and design of which is included in Appendix H.

Procedure

After participants read through the information contained on the Prolific entry on this study and elected to participate, they were provided a link to the main experiment. The main experiment of this study was housed on the Massey University's psychology lab server and consisted of three segments. It included an introduction to the experiment, the afore-mentioned measures and experimental tasks which made up the experimental task, and a post-experiment debrief. This section will describe the processes involved at each of these stages. A flow chart of the sequence of main events is included beneath in Figure 7.

Figure 7.

Sequence of the main events of the experiment.



A Google Slides presentation containing a visual facsimile of this experiment is also set up at: https://docs.google.com/presentation/d/1D-2DHme7SluBN5Fz2x_M7UVY50nXIiHbvRoNzwgXPqU. This series of slides contains an accurate demonstration of the overall design of the experiment, including the measures used along with the introductory and post-experiment debrief texts.

Main experiment

Introduction and consent

After clicking the link and arriving at the main experiment participants were greeted by a screen welcoming them to the study and asking them to carefully read through a couple information pages before deciding whether to take part. A copy of the information presented in these pages is included in Appendix I. After reading through this information participants were then presented with a consent form which asked participants to check a series of five boxes, indicating that they agreed with each statement if they were interested in participating. Participants were required to check each of these boxes before they could begin the experiment, as the continue button remained inactive until all of the boxes were ticked. A copy of the consent form is contained in Appendix J.

Measures and experimental tasks

After consenting to participate, participants were randomly presented with either the TEQ or MentS (with the additional attention check), as described above. After completing their initial measure, participants were directed to perform the measure they had yet to complete. After completing both the TEQ and MentS the participants were directed to complete the MCSDS-SF. Next, participants were presented with a series of instructions for the RMET-R section of the test and informed that on the upcoming screens they would be presented with a series of pictures of the eye regions of different individuals each paired with a variety of numbered mental state terms. Following this they were provided, and asked to, scan a list of the mental state terms used in the RMET-R and to look for and click on any that they were unfamiliar with to bring up a short definition and example of the term.

At this point, participants were randomly provided one of the three RMET-R variations (Long, Short, Occlusion). They were instructed depending on their variation how

long they would have to answer each question (Long: 20 sec, Short: 5 sec, Occluded: 5 sec), and in the Occluded variation how long the picture would be visible for (0.5 sec). Participants were also informed that if they failed to answer in time, they would receive a score of 0 for that question and be automatically taken to the next question. All participants then completed a practice question, with the following screen informing participants of the correct answer. Following this, participants completed their assigned variation of the RMET-R.

After completing the RMET-R's 36 questions the participants were then directed to answer the manipulation check questions and the second attention check. Following which they were directed to complete the RMET-R Methods Questionnaire designed by the authors of this experiment, with participants alternatively directed to type out and indicate what methods they used and how they used them to answer the RMET-R. After which followed the experiments Demographic Questionnaire, which participants were asked to complete.

Post-experiment debrief

After the completion of the experiment's Demographic questionnaire participants were then provided a short post-experiment debrief. The debrief included a brief information sheet, a couple of questions to assess for the presence of any errors in the study and what participants believed the study was about, and a couple of links to further information. A copy of this debrief is included in Appendix K. Finally, a link was provided at the bottom of the last page which brought participants back to Prolific, ending the study.

Chapter 3

Experiment results

This chapter describes the results of the main experiment described in the previous chapter and is made up of six main sections. The first two sections describe the exclusionary decisions of this study, and the assumption checks for the statistical analyses used in this study. The third section reports the results of a series of analyses based around the three manipulated versions of the RMET-R which make up the main IV of this study and describes the dummy coding performed to conduct the main regression analyses. This section includes a series of manipulation check analyses and analyses investigating potential demographic confounds. The fourth section reports a series of descriptive statistics and the results of a series of analyses based around the TEQ and MentS which make up the measured variables of this study. This section provides a preliminary analysis of the supposed reliability and validity of these measures in order to examine the appropriateness of their inclusion in this study, and to examine if they significantly differ between the three experimental conditions. The fifth section reports the results of the analyses used to investigate the main hypotheses of the study outlined in chapter one. The Final section reports the results of a qualitative data analysis performed to explore what strategies individuals use to complete the RMET-R. A series of exploratory analyses are also included in this section investigating if there is any relationship between the use of any specific strategy or amount of strategies and performance on the RMET-R, TEQ, or MentS. All of the quantitative analyses described in this chapter were conducted using frequentist null hypothesis statistical testing and performed using Jamovi version 1.0.

Data exclusions

Data from 30 participants, with 10 participants from the 'Long' condition, 13 from the 'Short' condition, and 7 from the Occluded condition was excluded from this analysis. Following the exclusionary criteria stated in the experiment's pre-registration we excluded thirteen participants datasets from analyses for having scores exceeding 2 SDs on the Marlowe-Crowne Social Desirability Scale – Short Form (MCSDS-SF); four participants for failing both attention checks in the experiment; one participant for answering less than 30% of the RMET-R's questions; and six participants for having TEQ scores outside 2.5 SDs and six others for having MentS scores outside 2.5 SDs. We also had one participant that had to be excluded due to a corrupted and incomplete dataset, and two participants that appeared to have completed the study twice – leaving near identical datasets that came from the same IP address. To remedy this, we selected to exclude the secondary set of both participants data. This leaving us with data from 351 participants with 113 in the long condition, 112 in the short and 126 in the occluded condition.

Data Assumption Checks

Assumption checks were carried out prior to analysis to determine the appropriateness of the planned parametric tests for analysing the data. The planned parametric tests to analyse the main hypotheses of this study included: (1) a one-way ANOVA to assess the effect of the three experimental manipulations on participants RMET-R scores. (2) Two hierarchal ordinary least squares (OLS) linear regressions to assess the effect of the experimental conditions on the relationship between (2a) RMET-R and TEQ scores, and (2b) RMET-R and MentS scores. (3) A Cronbach's alpha was calculated for each of the RMET-R conditions to assess the effect of the limitations on the test's internal consistency.

Several further analyses were also performed to conduct a series of manipulation and potential confound checks and to analyse a couple of secondary and exploratory aims of this

study. These analyses included a series of one-way ANOVAs, of Pearson product-moment correlations, Cronbach's alpha reliability coefficients, Spearman's rho correlations and Pearson's chi-square tests.

One-way ANOVA assumptions

For the one-way ANOVA the assumption of normality was assessed by visual inspection of Q-Q plots of the standardised residuals for each of the response variables, as shown in figures L1-L13 in appendix L. Despite the presence of some minor departures of normality on their ends the Q-Q plots of test difficulty, believed performance, RMET-R scores, RMET-R scores (occluded condition), RMET-R scores (long condition), TEQ scores, MentS scores, MentS Motivation scores, MentS Other scores and MentS Self scores all suggested that the assumption of normality was met for these variables.

The Q-Q plots of mental effort, RMET-R scores (short condition), and total strategies used all suggested that the data of these variables might be slightly skewed. To double check the normality of these variables we assessed their skewness and kurtosis, which is shown below in Table 2. This analysis showed that both mental effort and RMET-R scores (short condition) were moderately left skewed (Hair et al., 2010) and total number of strategies used was normally distributed. However as this departure of normality is not substantial, and the one-way ANOVA is robust against the violation of this assumption (Mena et al., 2017) no corrections were performed.

Table 2

Skewness and Kurtosis of mental effort, RMET-R scores (short condition) and Total number of strategies used

Variable	Skewness	Kurtosis
Mental effort	-0.667	0.126

RMET-R (short condition)	-0.408	0.049
Total number of strategies used	0.041	-0.566

The assumption of homogeneity of variances was assessed using Levene's Test for Homogeneity of Variances. Table 3 below displays that this assumption was met for mental effort required, RMET-R score (occluded condition), RMET-R score (short condition), RMET-R score (long condition), RMET-R score (Total), TEQ score, MentS score, MentS Motivation score, MentS Other score, and MentS Self score. However this assumption was not met for test difficulty, believed performance, or total number of strategies used. Due to this we ran a Welch one-way ANOVA to assess for differences in test difficulty, believed performance and total number of strategies used between the three RMET-R conditions, as the Welch one-way ANOVA does not assume equal variances.

Table 3

Results from Levene's Test of Equality of Variances for all one-way ANOVA variables

Variable	F	df1	df2	p
Test difficulty	4.867	2	348	0.008
Mental effort required	0.706	2	348	0.494
Believed performance	4.668	2	348	0.010
RMET-R score (occluded condition)	2.400	1	119	0.124
RMET-R score (short condition)	0.731	1	108	0.395
RMET-R score (long condition)	0.203	1	109	0.653
RMET-R score (Total)	0.629	2	348	0.534
TEQ score	1.147	2	348	0.319
MentS score	0.986	2	348	0.374
MentS Motivation score	0.095	2	348	0.909
MentS Other score	0.067	2	348	0.935
MentS Self score	0.470	2	348	0.625

Variable	F	df1	df2	p
Total number of strategies used	4.800	2	349	0.009

All possible means to make sure the assumption of independence was met were also performed. As the experiment assigned participants to a single experimental condition the only way this assumption could be breached was if a participant attempted to complete the assessment twice or got another participant in the same vicinity to perform the assessment. As we had little control over the participants we analysed for and excluded any datasets which appeared to be duplicates of another.

Regression analysis assumptions.

For our two linear regression analyses the analyses' assumptions of linearity and homoscedasticity were assessed by visual inspection of scatterplots of the predicted residuals, as shown in Figures M1 – M6 in Appendix M. Visual figures for the dummy coded variables were not included, as since the dummy variables of these regression are linear by default as they only involve two points (Hardy, 1993).

Both of these assumptions were confirmed in the TEQ-RMET-R regression with none of the scatterplots expressing any evidence of a nonlinear relationship or a breach of homoscedasticity (Hair et al., 2010). The residual plot of the MentS-RMET-R regression displayed little evidence of a breach of either of these two assumptions but presented with noticeable clumping amongst the datapoints of the X axis. This might indicate that there may be some issues related to the use of RMET-R scores to explain variation in MentS scores. This pattern suggests that the three subcategories which make up the MentS may vary considerably in their relationship with the RMET-R.

The assumption of multivariate normality was assessed via the calculation of Q-Q plots of the standardized residuals of each regression which are shown in figures N1 and N2

in Appendix N. These two plots show that the assumption of multivariate normality was met for both linear regressions. The assumption of no autocorrelation was assessed via the Durbin-Watson test. As displayed in Table 4 below all of the steps of the two regressions were found to have a Durbin-Watson statistic just below two, which indicates that there is no autocorrelation present in our two regressions.

Table 4

Durbin–Watson Test for Autocorrelation results for all steps of our hierarchal regression analyses

	Autocorrelation	DW Statistic	P
TEQ-RMET-R regression			
Step 1	0.0154	1.96	0.704
Step 2	0.0152	1.97	0.728
Step 3	0.0153	1.97	0.730
MentS-RMET-R regression			
Step 1	0.0022	1.98	0.834
Step 2	-0.0017	1.99	0.884
Step 3	-0.0012	1.99	0.906

The assumption of no, or little, multicollinearity was assessed by analysing the variance inflation factor for all of the variables in each step of our two regressions. As displayed below in Table 5 there was little multicollinearity in all variables in the first step of our two regressions. In step 2 and 3 there was a substantial rise in multicollinearity in both regressions, however this appears to be due to the inclusion of the interaction terms in step 2 and 3 and should not affect the overall regression. Due to this no corrections were performed.

Table 5

Collinearity statistics for all steps of our hierarchal regression analyses

	VIF	Tolerance
TEQ-RMET-R regression		
Step 1		
RMET Total score	1.25	0.803
Short (dummy)	1.53	0.653
Occluded (dummy)	1.65	0.607
Step 2		
RMET Total score	1.91	0.524
Short (dummy)	26.38	0.038
Occluded (dummy)	1.81	0.553
RMET Total score ✻ Short (dummy)	23.37	0.043
Step 3		
RMET Total score	4.40	0.227
Short (dummy)	40.47	0.025
Occluded (dummy)	40.95	0.024
RMET Total score ✻ Short (dummy)	33.98	0.029
RMET Total score ✻ Occluded (dummy)	32.34	0.031
MentS-RMET-R regression		
Step 1		
RMET Total score	1.25	0.803
Short (dummy)	1.53	0.653
Occluded (dummy)	1.65	0.607
Step 2		
RMET Total score	1.91	0.524

Short (dummy)	26.38	0.038
Occluded (dummy)	1.81	0.553
RMET Total score ✻ Short (dummy)	23.37	0.043
Step 3		
RMET Total score	4.40	0.227
Short (dummy)	40.47	0.025
Occluded (dummy)	40.95	0.024
RMET Total score ✻ Short (dummy)	33.98	0.029
RMET Total score ✻ Occluded (dummy)	32.34	0.031

Cronbach's alpha assumptions

Regarding the use of Cronbach's alpha, the RMET-R and TEQ both meet the assumption of Tau Equivalence and unidimensionality as both assessments were designed and have been shown to assess a single variable with the use of a single type of scale (Baron-Cohen et al., 2001a; Preti et al.,2017; Spreng et al., 2009). However, the MentS breaks the assumption of unidimensionality as although the test was designed to provide an overall measure of an individual's mentalizing abilities the assessment is actually made up of three separate subscales (Dimitrijević et al., 2018). This violation causes a major underestimate of reliability by the alpha, with Tavakol and Dennick (2011) suggesting that in cases such as this an alpha should be calculated for each the tests concepts or factors instead of the overall assessment. Therefore, in this study we will assess the Cronbach's alpha of all three subcategories of the MentS which on their own meet this assumption. Dimitrijević et al. (2018) partially followed this recommendation, as they generated Cronbach's alphas for the overall test and each of its subcategories. The nature of the participant recruitment in this

study makes the assumption of uncorrelated error terms very difficult to directly assess as we had little control over the participants. As we are not using time-series data and our experiment was open for individuals from multiple countries around the world this assumption is however likely met.

Pearson's R assumptions

Regarding the assumptions of Pearson's product moment correlation the assumption of level of measurement was met in all analyses as all the included variables are continuous. The assumption of absence of outliers was also met with all variable scores 1.5 interquartile ranges below the first quartile or above the third quartile removed before analysis. To ensure the assumption of related pairs was met before each analysis the variables were assessed to make sure there was a matching pair for every participant. Participants that did not have a matching pair of variables were removed from that analysis. The assumption of linearity was assessed by the visual inspection of scatterplots of the correlations variables which are shown in figures O1-O17 in Appendix O. The scatterplots show that this assumption was met with none of the scatterplots exhibiting any evidence of a curvilinear relationship.

Pearson's chi-square test assumptions

All of the assumptions of Pearson's chi-square test were met in this study, as all included variables (experimental condition, intuition, mental imagery, previous experience, rules of thumb, process of elimination, and multiple strategies) were independent and categorical in nature.

Spearman's rho assumptions

The assumption of ordinal, interval or ratio scale variables was met for this analysis as all the variables of interest are either dichotomous (dummy coded strategies) or continuous in nature (number of strategies used, RMET-R scores, etc).

Manipulated variable analyses and dummy coding

Manipulation checks

Prior to our main analysis we ran a series of manipulation check analyses to examine if our experimental time constraints effected the RMET-R in the manner that we theorised. As shown in Table 6 a preliminary analysis of the average scores of our three manipulation check questions found that participants responded in the manner we expected across the three experimental conditions. Participants that completed the occluded condition of the RMET-R rated it as more difficult, mentally effortful and reported that they had performed worse than participants who completed the short or long condition, while participants that completed the short condition of the RMET-R did the same compared to participants that completed the long condition of the RMET-R.

Table 6
Descriptive statistics of reported test difficulty, mental effort, and believed performance in each experimental condition

	Experimental condition	N	Mean	SD	SE
Test difficulty	Occluded	126	7.51	1.83	0.163
	Short	111	6.37	2.03	0.192
	Long	114	5.50	2.24	0.210
Mental effort	Occluded	126	7.60	1.89	0.168
	Short	111	7.16	1.97	0.187
	Long	114	6.44	2.05	0.192
Believed performance	Occluded	126	6.14	2.23	0.199
	Short	111	5.54	1.78	0.169
	Long	114	5.46	2.01	0.189

To examine if these differences were statistically significant we ran a series of one-way ANOVAs to examine the effect our three experimental conditions had on how participants rated the difficulty and mental effort of the test, and how well they believed they had performed.

A one-way ANOVA (Welch's) showed that RMET-R condition had a significant effect on how difficult participants rated the test ($F(2, 226) = 29.89$, $P < .001$). Post hoc testing using the Games-Howell test found that participants in the occluded condition rated the test as significantly more difficult than participants in the short ($P<.001$) and long condition ($P<.001$). Participants in the short condition rated the test significantly more difficult than those in the long condition ($P=.007$).

A one-way ANOVA showed that RMET-R condition also had a significant effect on how mentally effortful participants rated the test ($F(2, 348) = 10.6$, $P < .001$). Post hoc testing using Tukey's correction found that participants who completed the occluded condition rated the RMET-R as significantly more mentally effortful than participants who completed the long condition ($P<.001$), but not significantly different than individuals who completed the short condition ($P=.198$). Participants who completed the short condition rated the RMET-R as significantly more mentally effortful than participants who completed the long condition ($P=.017$).

A one-way ANOVA(Welch's) also showed that RMET-R condition had a significant effect on how well participants believed they had performed on the RMET-R ($F(2, 232) = 3.75$, $P = .025$). Post hoc testing using the Games-Howell test found that participants who performed the occluded condition believed they had performed significantly worse on the RMET-R than individuals who completed the long condition ($P=.034$) but no significantly different than individuals who completed the short condition ($P=.057$). Contrary to as expected individuals who completed the short condition did not believe they performed any significantly different on the RMET-R then individuals in the long condition ($P=.941$).

These patterns of results suggest that our experimental constraints affected the difficulty and mental effort of RMET-R in the manner that we theorised. The lack of a

significant difference between our occluded and short conditions in how participants rated the mental effort of the test may suggest that the effect of adding stimulus time constraints might not be quite as strong as we originally believed though. The effect of the three constraints on how participants believed they performed is also less than expected, though this may possibly be due to the design of this question, which in retrospect likely encourages neutral responding. In any case the results of these manipulation checks are more than enough to support the conduction of our planned analyses to assess the main hypotheses of this study.

Analyses for potential confounding variables

Prior to our main analysis we also ran a series of analyses to check if a couple of demographic variables were confounded with our IV manipulation. Specifically we were interested in assessing if participants' age, gender, or education level could have a confounding effect on RMET-R performance between the three levels of our independent variable. As previously mentioned these three variables have been found to effect RMET-R performance in previous research with older, male, and poorer educated individuals scoring more poorly on the test. Theoretically these individuals are likely to be more strongly affected by our experimental constraints, which could easily affect the validity of our planned regressions results, especially if their distribution differs between the three conditions.

To assess if these three demographic variables had a confounding effect on RMET-R scores we conducted a series of Pearson product-moment correlations to investigate whether participant age and education level had a significant effect on RMET-R scores across the experiments three conditions. We also conducted a series of one-way ANOVAs to assess if participant gender had a significant effect on RMET-R scores in the experiment's three conditions. All participants' datasets that reported that they were gender diverse were

removed from this analysis, due to the very small number of individuals that made up this category.

As displayed in Table 7 neither participant age nor participant level of education were found to be significantly correlated with RMET-R score in any of the three experimental conditions. Likewise as displayed in Table 8 a series of three one-way ANOVA's showed that participants gender had no significant effect on RMET-R scores in each of the three experimental conditions. This pattern of results shows that participants age, gender, and education level is not related to and has no confounding effect on RMET-R performance.

Table 7

Descriptive statistics and correlations between participant age, education level and RMET-R scores

Variable	M	SD	RMET-R correlation (r)	*P*
Participant age				
Occluded condition	29.70	7.53	-0.089	0.324
Short condition	27.70	7.20	-0.022	0.822
Long condition	29.50	7.74	-0.153	0.105
Participant education level				
Occluded condition	5.21	1.54	-0.037	0.934
Short condition	4.96	1.63	-0.062	0.518
Long condition	5.15	1.58	0.059	0.530

Table 8

Means, standard deviations, and one-way analysis of variance of participant gender (Male/Female) and RMET-R scores by experimental condition.

Experimental condition	Female		Male		ANOVA		
	M	SD	M	SD	*F* ratio	DF	*P*
Occluded condition							

RMET-R score	21.6	4.26	20.1	5.16	3.01	2, 119	0.085
Short condition							
RMET-R score	22.8	4.40	21.6	5.28	1.89	1, 108	0.172
Long condition							
RMET-R score	26.5	4.26	26.1	4.42	0.24	1, 109	0.622

Main regression analysis dummy coding

Prior to the main analysis the three experimental conditions (Long, Short, Occluded) that made up the main independent variables (IV) of this study's main hypotheses were transformed into two dummy variables for the linear regression. A Short dummy variable (coded 1 for Short condition, and 0 for not Short condition) and an Occluded dummy variable (coded 1 for Occluded condition, and 0 for not Occluded condition) were created in the dataset. The Long category used as the reference category (represented by 0 on both dummy variables). This coding allowing us to assess for differences between the Long and Short and Long and Occluded conditions in our planned regression.

Measured Variables analyses

Before conducting our analyses to examine our main hypotheses we conducted a series of analyses to examine the validity and reliability of the TEQ and MentS to provide a preliminary analysis of the appropriateness of their inclusion. To this end we conducted a series of Pearson product-moment correlations to examine if the TEQ, MentS and its subscales positively correlated with each other as they theoretically should, given the close theoretical relationship between empathy, mentalizing, and ToM (Spreng et al., 2009; Dimitrijević et al., 2018). To examine the reliability of the TEQ and MentS we conducted a series of a Cronbach's alpha reliability coefficients. As the MentS violates the analyses assumptions of Tau Equivalence and unidimensionality we will instead analyse the internal

consistency of the three subscales that make up the MentS and examine how well the three subscales correlate with each other. If the MentS is a reliable measure of overall mentalizing ability as claimed by Dimitrijević et al. (2018) all of its subscales should demonstrate good internal consistency and be correlated with each other. Dimitrijević et al. (2018) reports that all of the subcategories are positively correlated with each other.

To assess if there were any significant differences in TEQ, MentS, MentS Motivation, MentS Other, or MentS Self scores between the three experimental conditions we also ran a series of one-way ANOVAs. A significant difference in the scores of these variables between the three experimental conditions could adversely impact the validity of our planned regression analyses results.

Toronto Empathy Questionnaire

As displayed in Table 9 a series of Pearson product-moment correlations showed that the TEQ was positively correlated with the MentS and the MentS's Motivation and Other subscales, and negatively correlated with Self subcategory of the MentS. This pattern of results fits exactly in the manner it theoretically should, with a particularly strong positive relationship found between the TEQ and the Other subscale of the MentS which supports the supposed concurrent validity of the test. The negative relationship between the TEQ and the Self subscale of the MentS though not predicted is not outside of expectations as these two concepts have no theoretical reason to be related to each other. A Cronbach's alpha reliability coefficient showed that the TEQ had a good level of internal consistency, $\alpha = .881$, with all of the assessments questions positively correlated with each other to a good degree (lowest $r = .375$), which indicates that the TEQ has a good level of reliability. A one-way ANOVA showed that there were no significant differences in TEQ scores between the three conditions ($F(2, 348) = 0.455$, $P = .635$). This result indicates that there is no variance in TEQ scores between the three conditions which could impact the findings of our regression analysis.

Mentalizing Questionnaire

As described in the previous section the results of our Pearson product-moment correlations supported the supposed validity of the MentS with the overall MentS and its Motivation and Other subscales positively correlated with the TEQ. However as displayed in Table 9 the Self subcategory is negatively correlated with the Other subcategory and has a just beneath statistical significance (P = 0.51) negative correlation with the Motivation subcategory of the MentS. This pattern of results is significantly different to the pattern of results reported by Dimitrijević et al. (2018) and suggests that the MentS has poor internal consistency.

Table 9

Descriptive statistics and correlations for TEQ, MentS and MentS subscales

Variable	M	SD	1	2	3	4	5
1 TEQ	45.4	8.19	—				
2 MentS	93.6	8.04	0.360***	—			
3 MentS Motivation	34.8	3.85	0.364***	0.715***	—		
4 MentS Other	37.6	4.94	0.595***	0.584***	0.510***	—	
5 MentS Self	21.2	5.52	-0.269***	0.411***	-0.100	-0.354***	—

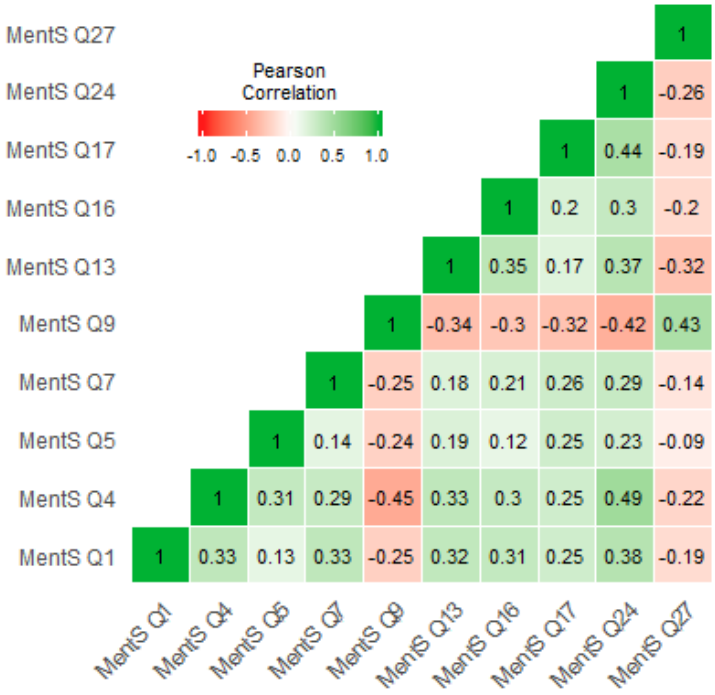
Note. *** p < .001

Our analysis of the internal consistency of the MentS's subcategories showed that the Self subcategory of the MentS had a good level of internal consistency, α = .777, with all of the subcategory's questions positively correlated with each other (lowest r = .307). The Other subcategory had a good level of internal consistency, α = .810, with all of the

subcategory's questions positively correlated with each other (lowest r = .272). However, the MentS Motivation subcategory had a poor level of internal consistency, α = .467. As displayed in Figure 8, two of the questions which make up the subscale are negatively correlated with the rest of the items in the scale. This pattern of results is quite different to the results reported in Dimitrijević et al. (2018) and raises a number of questions about the supposed reliability of the assessment. Our results suggest that although the MentS and its Other subcategory displays good construct validity the overall test has poor internal consistency and doesn't fit its supposed three factor design well.

Figure 8

Correlation heatmap of the MentS Motivation subscale of the MentS



A series of one-way ANOVA showed that between the three experimental conditions there were no significant differences in MentS scores (F (2, 348) =0.919, P = .400), MentS Motivation scores (F (2, 348) =1.131, P = .324), MentS Other scores (F (2, 348) =0.926, P = .397), or MentS Self scores (F (2, 348) =0.356, P = .701). This result indicates that there is

no variance in MentS or its subscales scores between the three conditions which could impact the findings of our regression analysis.

Main analyses

This section describes the results of the analyses performed to test the main hypothesis of this study outlined in the previous chapter. These analyses are divided into three sections, with the first assessing if our experimental time constraints increased the difficulty of the RMET-R, the second assessing if they increased the tests concurrent validity, and the third assessing if they increased the internal consistency of the test in the manner that we hypothesised.

Difficulty

To investigate if our experimental constraints increased the difficulty of the RMET-R we first conducted a one-way ANOVA to assess if RMET-R scores significantly differed between our three experimental manipulations. A one-way ANOVA showed that our experimental manipulations had a significant effect on RMET-R score ($F(2, 348) = 42.0$, $P < .001$). Table 10 displays the mean RMET-R scores in each of the study's three experimental conditions. Post hoc testing using Tukey's correction found that participants scored significantly lower in the occluded RMET-R condition compared to the long condition ($P<.001$), and in the short condition compared to the long condition ($P<.001$). However, there was no significant difference in participants scores between the occluded and short conditions of the RMET-R ($P= .170$).

This pattern of results indicates that as we predicted adding increased answer time constraints to the RMET-R increases the difficulty of the assessment. These results also indicate that although adding stimulus time constraints to the RMET-R increases the difficulty of the assessment, it is at a notably lower rate than we theorised.

Table 10

Descriptive statistics of RMET-R score in each experimental condition

	N	Mean	SD	SE
RMET-R score				
Occluded condition	126	21.1	4.69	0.418
Short condition	111	22.2	4.90	0.465
Long condition	114	26.3	4.31	0.404

Validity

This section describes the results of the two hierarchal multiple linear regressions we conducted to assess our pre-registered hypotheses. Therefore this section is divided into two sections, with the first assessing if our experimental constraints improved the relationship between the RMET-R and the TEQ and the second assessing if they improved the relationship between the RMET-R and the MentS.

TEQ-MentS regression. To examine if our experimental constraints improved the relationship between the RMET-R and the TEQ we conducted a three-step linear regression with TEQ scores as the dependant variable. All of the steps and variables included in this regression along with its results are displayed below in Table 11. The first step of the model was used to investigate our first hypothesis that RMET-R scores would be positively related with TEQ scores. The results of this first model indicate that this hypothesis was met with the RMET-R found to significantly predict TEQ scores when controlling for condition. The interaction effect in the second step of the model was used to test our second hypothesis that RMET-R scores would more strongly covary with TEQ scores when RMET-R answer time limits are shorter. The results of our analysis indicated that this hypothesis was not met, as the interaction effect between RMET-R scores and the short condition and the $\Delta R2$ between the first and second model of the regression was not significant. These results show that there is

next to no difference in the relationship between the RMET-R and the TEQ between our short and long conditions of the assessment.

The interaction effect in the third step of the model was used to test our third hypothesis that RMET-R scores would more strongly covary with TEQ scores when the time of the RMET-Rs stimuli was shorter. Akin to the second hypothesis the results of this analysis indicated that this hypothesis was not met, with a negligible difference in the relationship between the RMET-R and the TEQ existing between the occluded condition of the RMET-R and our other two conditions. No significant interaction was found between RMET-R scores and the occluded condition, and no significant ΔR2 between the second and third model of the regression. As both of these interaction effects were not statistically significant we did not perform any follow up simple slopes analyses to further investigate how our experimental manipulations effected the relationship between the RMET-R and TEQ.

Table 11

Hierarchal regression results for TEQ-RMET-R regression

Variable	B	95% CI for B		SE B	B	R^2	ΔR^2
		Lower	Upper				
Step 1						0.0181	
Intercept	38.949***	33.823	44.074	2.606			
RMET-R score	0.221***	0.034	0.407	0.095	0.139		
Short (dummy)	1.854	-0.431	4.140	1.162	0.105		
Occluded (dummy)	2.041	-0.253	4.335	1.166	0.120		
Step 2						0.0181	2.13e-5
Intercept	39.106***	32.849	45.363	3.181			
RMET-R score	0.215	-0.016	0.445	0.117	0.138		
Short (dummy)	1.449	-8.054	10.951	4.831	0.105		
Occluded (dummy)	2.010	-0.396	4.415	1.223	0.118		

RMET Total score ✻ Short (dummy)	0.017	-0.375	0.409	0.199	0.005		
Step 3						0.0181	1.04e-5
Intercept	39.319***	29.958	48.680	4.759			
RMET-R score	0.207	-0.145	0.558	0.178	0.138		
Short (dummy)	1.235	-10.551	13.022	5.992	0.103		
Occluded (dummy)	1.666	-9.801	13.132	5.830	0.117		
RMET Total score ✻ Short (dummy)	0.025	-0.448	0.498	0.240	0.007		
RMET Total score ✻ Occluded (dummy)	0.014	-0.452	0.481	0.237	0.004		

Note. * p < .05, ** p < .01, *** p < .001

MentS-RMET-R regression. To examine if our experimental constraints improved the relationship between the RMET-R and the MentS we conducted a three-step linear regression with MentS scores as the dependant variable. All of the steps and variables included in this regression along with its results are displayed below in Table 12.

The first step of the model was used to investigate our first hypothesis that RMET-R scores would be positively related with MentS scores. The results of this first model indicate that this hypothesis was not met as it was found that the RMET-R did not significantly predict MentS scores when controlling for condition. The interaction effect in the second step of the model was used to test our second hypothesis that RMET-R scores would more strongly covary with MentS scores when RMET-R answer time limits are shorter. The results of our analysis indicated that this hypothesis was not met, as the interaction effect between RMET-R scores and the short condition and the ΔR2 between the first and second model of the regression is not significant. These results indicate that there is next to no difference in

the relationship between the RMET-R and the MentS between our short and long conditions of the assessment.

The interaction effect in the third step of the model was used to test our third hypothesis that RMET-R scores would more strongly covary with MentS scores when the time of the RMET-Rs stimuli was shorter. The results of this analysis indicated that this hypothesis was also not met, with a negligible difference in the relationship between the RMET-R and the MentS existing between the occluded condition of the RMET-R and our other two conditions. No significant interaction was found between RMET-R scores and the occluded condition, and no significant ΔR2 between the second and third model of the regression was found. As both of these interaction effects were not statistically significant no follow up simple slopes analyses to investigate how our experimental conditions effected the relationship between the RMET-R and MentS were performed.

Table 12

Hierarchal regression results for MentS-RMET-R regression

Variable	B	95% CI for B		SE B	B	R^2	ΔR^2
		Lower	Upper				
Step 1						0.0055	
Intercept	92.687***	87.623	97.750	2.574			
RMET-R score	0.013	-0.171	0.197	0.094	0.008		
Short (dummy)	1.498	-0.761	3.754	1.148	0.086		
Occluded (dummy)	0.475	-1.791	2.740	1.152	0.028		
Step 2						0.0099	0.0043
Intercept	94.888***	88.721	101.056	3.136			
RMET-R score	-0.071	-0.298	0.157	0.116	0.003		
Short (dummy)	-4.178	-13.545	5.189	4.762	0.080		

Occluded (dummy)	0.035	-2.337	2.406	1.206	0.002		
RMET Total score ✻ Short (dummy)	0.241	-0.145	0.627	0.196	0.072		
Step 3						0.0164	0.0066
Intercept	100.149***	90.952	109.346	4.676			
RMET-R score	-0.271	-0.616	0.074	0.175	-0.003		
Short (dummy)	-9.439	-21.019	2.141	5.888	0.043		
Occluded (dummy)	-8.445	-19.711	2.820	5.728	-0.018		
RMET Total score ✻ Short (dummy)	0.441	-0.024	0.906	0.236	0.131		
RMET Total score ✻ Occluded (dummy)	0.353	-0.105	0.811	0.233	0.108		

Note. * p < .05, ** p < .01, *** p < .001

This pattern of results indicates that our hypotheses were not met, with our experimental answer time and stimulus limitations not significantly effecting the concurrent validity of the RMET-R. Contrary to as we expected these results also showed that in our sample RMET-R scores did not covary with MentS scores in the manner that they theoretically should, with no significant relationship existing between these two variables. In combination with the poor internal consistency found with the MentS earlier this suggests that there are some psychometric issues with the assessment. Although TEQ scores covaried with RMET-R scores in the manner we hypothesized the relationship between them was smaller than we expected, and its significance disappeared when we included the interaction terms in the model in step two and three. This suggests that although there is a positive relationship between TEQ and RMET-R scores in this sample it is less than we theorised, which might be due to the neurotypical sample we used in this experiment.

As MentS scores were found to be not significantly related to RMET-R scores we conducted a quick Pearson product-moment correlation to assess if MentS Other scores could be used as an alternative dependant variable in the same type of regression, as it should theoretically be strongly related to RMET-R scores. However the analysis showed that MentS Other scores were not correlated with RMET-R scores in our sample $r(342) = .030$, $p = .582$. This result indicates that there is no value in conducting a third regression and suggests that the MentS has poor construct validity.

Reliability

To assess if our experimental time constraints increased the internal consistency of the RMET-R as we hypothesised we calculated a Cronbach's alpha for each of our three RMET-R conditions. As displayed in Table 13 we found that in contrast to as we hypothesised the internal consistency differed little between our three conditions with a negligible increase in the short condition and a negligible decrease in the occluded condition compared to the long condition, with all conditions shown to have a questionable level of internal consistency.

Table 13

Cronbach's alpha results for each RMET-R condition

RMET-R condition	Cronbach's Alpha
Occluded condition	0.651
Short condition	0.689
Long condition	0.658

This lower level of reliability is a noted issue with the RMET-R and has been reported before in previous research using the standard version of the assessment (Ferguson & Austin, 2010; Hayward & Homer, 2017) which makes it unlikely that this low level of reliability is due to our experimental manipulations. These results therefore indicate that our constraints

do not increase the reliability of the RMET-R as we hypothesised, with our constraints instead having no major impact on the tests reliability.

Strategies used to complete the RMET-R

As discussed in the previous chapters there has been no major research conducted which has looked into how the RMET-R is actually performed by its participants. Due to this lack of information we selected to include a single open-ended survey question situated after the RMET-R to allow us to conduct a small qualitative analysis to investigate what strategies they reported using. A qualitative analysis was chosen to be performed as it allowed us to conduct a much wider and more open-ended exploration than a quantitative analysis would allow and allowed us to avoid influencing the participants responses (Rahman, 2017). This gave us a sizable amount of data which we then analysed and coded to identify patterns and themes which, given the comparatively small scale of data for a qualitative analysis, was done by hand. The coding of our data followed grounded theory and open coding principles and was partially data and theory driven (Strauss & Corbin, 1990; Kawulich, 2004). Prior to the comment of the experiment we came up with five strategies that we theorised were used based off the results of some preliminary pilot study data and prior research on how affective ToM and the RMET-R function.

Our analysis of the data found that most participants reported the use of strategies that reliably fit into one of our five priorly theorised strategies, with only a minor percentage of individuals reporting the use of an alternative strategy. In total our analysis identified the use of eight major strategies that were used by participants to complete the RMET-R. Table 14 below displays these main strategies along with an example quote from various participants and the number of times they were used across the study. A small number of participants were also found to report using multiple strategies as their main means of completing the

assessment, with their answers fitting into more than one category. These individuals were placed in a category of their own with the combinations of methods that this subcategory of participants reported using recorded in Table 15.

Table 14

Main strategies reported by participants that were used to complete the RMET-R across the entire experiment (N = 351)

Main strategy	Example quote	Frequency , n (%)
1: Intuition (Initial impression/no strategy)	"I just went by my initial impression. I chose the word that best described the emotion I perceived."	67 (19.3%)
2: Mental imagery (Imagining the rest of the face/ environment)	"Tried to imagine the rest of the face or a situation that the look would occur in"	27 (7.8%)
3: Previous experience (Comparing the eyes to previous experiences /events)	"my method looked at picking the word that describes the eyes shown with my own experience of those feelings and the way people have expressed these feelings; so if the eyes were similar to expressions I had seen before and the feelings that were experienced with it"	39 (11.2%)
4: Rules of thumb (eye direction, level of squint, brow shape)	"I think I mostly went by the direction the eyes were pointing (whether towards or away from the viewer) and the position of the eyebrows (furrowed; relaxed; lifted; etc.). These small expressions help me read a face."	167 (48%)
5: Process of elimination (Picked best answer/ ruled out through previous knowledge)	"Quick look at the eyes; then the words; narrowed it down to 2 then picked what I thought was the correct one"	26 (7.5%)
6: Mimicry (copying look of the eyes)	"I tried to mimic the expression in the images. This felt like it allowed me to simulate the feeling."	5 (1.4%)
7: Putting self into mindset	"I tried to put myself in their shoes to understand the emotions"	1 (0.3%)
8: Empathetic response to picture	"I thought about how I would feel if they were looking at me."	1 (0.3%)

9: Combination of methods	"I looked at the picture and got an impression of the feeling; based on direction of their gaze and visible tension/wrinkling around the eyes; then looked at the words around the picture to see which most closely matched the impression I got." (combination of rules of thumb and process of elimination)	15 (4.3%)

Table 15

Combinations of methods reported by individuals using multiple methods to complete the RMET-R

Combinations of methods	**Number of participants (specific strategies combination)**
Mental Imagery and Rules of thumb	6
Previous experience and Rules of thumb	4
Rules of thumb and Process of elimination	2
Mental Imagery and Previous experience	1
Intuition and Rules of thumb	1
Intuition and Process of elimination	1

In our analysis the most commonly reported main strategy by a considerable margin was the use of rules of thumb, with nearly half of all the studies participant reporting they based their answers off a number of simple, generalised rules based around what specific facial features were displayed by the actors in the RMET-R's pictures. Participants commonly remarked that they based their answers on the shape of the actor's brow or the direction their eyes were looking. The next most commonly used strategy was the use of participants intuition or a lack of a strategy, with just under 20% of the overall sample reporting this as their main strategy. The third most commonly used strategy was the use of individuals' previous experiences to work out what the eyes of the actors were displaying, with a number of participants remarking that they based their answers off their recollections

of what other individuals or their own eyes looked like when expressing the mental states which made up the tests answers. The last two theorised strategies were reportedly to be used at around the same rate with the use of mental imagery to visualise the rest of the actor's face, and the use of knowledge about the mental state terms to guess the best answer both used by just above 7% percent of participants. The last three new strategies were found to only be used by an exceedingly minor number of participants, with only a single individual reporting the use of empathetic abilities and attempting to put themselves into the mindset of the individual as their main means of completing the assessment. The last strategy that was reported was rather unusual, with five individuals claiming that they mimicked the face that the actors in the test were displaying to work out the answers of the RMET-R. The use of this strategy by multiple participants is surprising as the feasibility of this method is highly suspect, especially in the shorter two conditions as it would require a reflective surface and take a fair bit of time to use.

To examine if the addition of our experimental time constraints significantly affected how individuals performed the RMET-R we ran a series of Pearson's chi-square tests to assess if the rate individuals used these main strategies significantly differed between the three experimental conditions. To do this a selection of dummy codes were created for the use of intuition, mental imagery, previous experience, rules of thumb, process of elimination, and multiple strategies (coded 0 for not this method, and 1 for this method). Due to the low number of participants that reported using mimicry, putting themselves into the mindset of the eyes, and their empathetic response as their main means of completing the RMET-R these methods were not included in the analysis.

The series of chi-square test showed that that there was no significant difference between the mental imagery X^2 (2, N = 348) = 3.75, p = .154, previous experience X^2 (2, N = 348) = 0.997, p = .607, process of elimination X^2 (2, N = 348) = 0.103, p = .950, and multiple

strategies X^2 (2, N = 348) = 2.28, p = .320 as their main strategy to complete the RMET-R. There was also a significant difference between the three experimental conditions in the percentage of participants that used Intuition X^2 (2, N = 348) = 12.4, p = .002 and rules of thumb X^2 (2, N = 348) = 10.1, p = .006 as their main strategy to complete the RMET-R. As displayed in Table 16 participants used intuition significantly more and rules of thumb significantly less as their main strategy in the occluded condition compared to the short and long condition. This pattern of results suggests that our addition of visual stimuli time constraints may have disrupted individuals' abilities to use rules of thumb, potentially forcing individuals to use quicker intuitive or guess based techniques.

Table 16
Frequency table of main strategies used to complete the RMET-R that were reported by participants in each experimental condition (N = 351)

Strategy	**Long condition**	**Short condition**	**Occluded condition**
1: Intuition	18 (15.8%)	13 (11.8%)	36 (29%)
2: Mental Imagery	13(11.4%)	5(4.5%)	9(7.3%)
3: Previous experience	12(10.5%)	15(13.6%)	12(9.7%)
4: Rules of thumb	58(50.9%)	63(57.3%)	46(37.1%)
5: Process of elimination	8(7.0%)	8(7.3%)	10(8.1%)
6: Mimicry	1(0.9%)	2(1.8%)	2(1.6%)
7: Putting self into mindset	0	0	1(0.8%)
8: Empathetic response to picture	1(0.9%)	0	0
9: Combination of main methods	3(2.6%)	4(3.6%)	8(6.5%)

As we were interested in assessing how many strategies participants reported using we created a new variable 'amount of methods used' by combining the number of main strategies with the number of secondary strategies participants reported using. Upon analysis we found that the majority of individuals reported the use of multiple strategies to answer the RMET-R, with most using a number of secondary strategies. In total 87.5% of participants

reported that they used more than one method to complete the RMET-R, with participants on average using three strategies.

As displayed in Table 17 the total number of strategies used was found to differ slightly between the three conditions. The long condition displayed a reasonably normal distribution in the number of strategies used, while participants in the short condition strongly favoured the use of three strategies, and participants in the occluded condition typically used more strategies. To investigate if the number of strategies used by participants significantly differed between our three conditions we ran a one-way ANOVA to assess if the total number of strategies used differed between the three experimental conditions. A one-way ANOVA test (Welch's) showed that there was a significant difference in how many strategies participants used between the three experimental conditions intuition ($F(2, 232) = 6.28$, $P = .002$). Post hoc testing using the Games-Howell test found that participants used significantly less strategies in the occluded condition than in the short condition ($P = .003$) and long condition ($P = .019$), but no significance differences between the short and long conditions ($P = .896$). This may be due to the more difficult and complex nature of the manipulation of the RMET-R in this condition, which may discourage the use of multiple strategies.

Table 17
Frequency table of numbers of strategies used by participants in each experimental condition and the total experiment.

	Experiment condition			
Total strategies used	**Long**	**Short**	**Occluded**	**Total**
1	10(8.8%)	10(8.9%)	24(19%)	44(12.5%)
2	32(28.1%)	18(16.1%)	39(31%)	89(25.3%)
3	38(33.3%)	54(48.2%)	38(30.2%)	130(36.9%)
4	25(21.9%)	24(21.4%)	21(16.7%)	70(19.9%)
5	9(7.9%)	6(5.4%)	4(3.2%)	19(5.4%)

To analyse whether there is any relationships between the use of any specific main, or quantity of, strategies and RMET-R, TEQ, or MentS scores a series of Spearman's rank correlation coefficients were performed using the dummy codes mentioned earlier. The three newly discerned strategies of mimicry, placing oneself into the mindset of the actor, and relying on empathetic response were not included in this analysis due to the small number of participants which reported using each strategy. Individuals using a combination of strategies were also not included either due to the small number of individuals reporting the use of multiple strategies and the combinatory nature of this category.

The analyses showed no significant correlations between any of the main strategies and MentS scores across any of the experiment's conditions. Likewise, there was no significant correlation between the use of any main strategies and TEQ and RMET-R scores with the exception of intuition. A significant negative correlation was discovered between the use of intuition and TEQ scores in the short condition, but the lack of a significant effect in either of the other conditions or across the experiment suggests this is due to noise. Likewise, a negative trend was found between the use of intuition and RMET-R scores. This trend was just below statistical significance in the short (P = .061) and occluded (P = .072) conditions, but statistically significant across the entire dataset. Table 18 displays the results of all of the correlations between the 5 strategies and the RMET-R, TEQ, and MentS across the entire experiment and in each of the experimental conditions.

Table 18

Correlation table (Spearman's rho) of the use of intuition, mental Imagery, previous experiences, rules of thumb and process of elimination with RMET-R, TEQ, and MentS scores across the entire experiment and in each experimental condition.

	Intuition	Mental Imagery	Previous experiences	Rules of thumb	Process of elimination

RMET-R					
Long	-0.106	0.105	-0.091	0.049	0.029
Short	-0.179	-0.046	0.049	0.069	0.100
Occluded	-0.162	-0.010	0.044	0.023	0.134
Total	-0.182***	0.057	-0.005	0.078	0.090
TEQ					
Long	0.137	0.137	0.137	0.137	0.137
Short	-0.035*	-0.035	-0.035	-0.035	-0.035
Occluded	-0.138	-0.138	-0.138	-0.138	-0.138
Total	-0.023	-0.023	-0.023	-0.023	-0.023
MentS					
Long	0.015	0.015	0.015	0.015	0.015
Short	-0.010	-0.010	-0.010	-0.010	-0.010
Occluded	0.006	0.006	0.006	0.006	0.006
Total	-0.029	-0.029	-0.029	-0.029	-0.029

Note. * $p < .05$, *** $p < .001$

As displayed in Table 19 the analysis showed that there were no significant correlations between the total number of strategies used and TEQ scores in any of the experimental conditions. However, a couple of significant correlations were found between the number of strategies used and MentS and RMET-R scores. A positive relationship was found to exist between the number of strategies used and the MentS in the long condition and in the total dataset. A positive relationship was also found to exist between the number of strategies and RMET-R scores in the short condition and across the total dataset.

Table 19

Correlations (Spearman's rho) between number of strategies used and TEQ, MentS and RMET-R scores across the entire experiment and in each experimental condition.

	TEQ	MentS	RMET-R
Total number of strategies used (long condition)	0.015	0.196*	0.065
Total number of strategies used (short condition)	0.134	0.183	0.227*
Total number of strategies used (occluded condition)	0.130	0.104	0.123
Total number of strategies used (Total study)	0.083	0.160**	0.174***

Note. * $p < .05$, ** $p < .01$, *** $p < .001$

Chapter 4

Discussion

This chapter provides a discussion of the results of the previous chapter, going over the meaning and relevance of our results as they relate to the aims and hypotheses of our study and the theoretical and practical significance of our findings to the wider research community. This chapter begins with a brief summary of the main and secondary aims, along with the corresponding findings of our study. Following this we discuss the implications of our study for the RMET-R, the other measures involved in our study, other ToM assessments, ToM compensation, and ToM researchers. The strengths and limitations of the current study are acknowledged, along with suggestions to fix issues with this study and recommendations for future research. This chapter ends with a reflection on the conclusions of this study and its contributions to the existing literature.

Summary of study aims and key findings

Recent research has suggested that many ToM assessments have a range of psychometric issues and can be easily passed using compensation, completely invalidating the assessment (Canty, 2016; Livingston & Happé, 2017). One method that might improve available ToM assessments and help deny the use of compensatory strategies may lie in the use of increased answer and stimuli presentation time constraints. Therefore, the primary aim of this study was to investigate if shortening the answer or stimulus presentation time could improve a ToM assessment's difficulty, validity, and reliability. The RMET-R was selected due to its high level of popularity and because it has a number of psychometric issues which theoretically make it easy to pass via a number of compensatory strategies. As a secondary exploratory aim of this study we were also interested in investigating how individuals completed the RMET-R as this has not been investigated before. In this study we were specifically interested in assessing if our experimental constraints effected how individuals

completed the assessment or if there was any relationship between the use of any specific, or quantity of, methods and RMET-R, TEQ or MentS scores.

The results from our analyses indicated that our experimental answer time constraints significantly increased the difficulty of the RMET-R in the manner that we theorised. However, although our results indicated that our experimental stimuli presentation time constraints increased the difficulty of the assessment it was not at a statistically significant level. This may suggest that neurotypical individuals affective ToM abilities are extremely fast, and that this form of constraint is less effective than theorised, as discussed later in this chapter.

Contrary to as we hypothesised the results of our experiment indicated that our experimental answer and stimulus time constraints had no effect on the validity of the RMET-R. The results of our analysis found that our three experimental conditions did not moderate the relationship between RMET-R and TEQ scores. Unexpectedly we found that there was no significant relationship between RMET-R and MentS scores, even though these measures should theoretically be related to one another. Combined with the results of a number of preliminary analyses this suggests that the MentS has a number of psychometric issues, as discussed later in this chapter. In addition, although our results indicated that the TEQ was related to the RMET-R in the manner that it theoretically should be the effect size was smaller than expected. Furthermore, its relationship lost its significance when we added the two interaction terms into the model, suggesting that the relationship between these two measures was not a strong as we originally believed. We also conducted a series of preliminary assessments to check for the presence of a number of potential confounding variables. The results of our analysis showed that there were no significant differences in participants' gender, TEQ scores, MentS scores, MentS Motivation scores, MentS Other scores, and MentS Self scores between our three experimental conditions and that RMET-R

scores were not significantly affected by participants' age or level of education between the three experimental conditions. Also counter to as hypothesised, the results from this experiment found that there were no major differences in reliability between our three variations of the RMET-R. A series of Cronbach's alphas demonstrated that our three variations all had a questionable level of internal consistency.

The results of our qualitative analysis into what strategies individuals used to complete the RMET-R found that individuals reported the use of eight strategies, with five major strategies reported by participants. In order of most to least used strategies they were: Rules of thumb, Intuition, Previous experiences, Mental Imagery, and Process of elimination. A number of individuals reported that they used a combination of these methods as their main means of completing the RMET-R. The remaining three strategies used to complete the RMET-R were used by a minor ($N \leq 5$) number of participants. They were, in order from most to least used; Mimicry, Putting oneself into the mindset of the stimuli's actor, and Empathetic response to picture.

The results of this study found that there were no significant differences in the use of previous experiences, mental Imagery, and process of elimination across our three experimental conditions. However, in the occluded condition participants reported using Rules of thumb significantly less and intuition significantly more than in the other two conditions. These results suggest that our stimuli time constraints might affect individuals' abilities to use the mostly commonly used means to complete the RMET-R, as discussed later in this chapter. All the most commonly used strategies were found to have no relation with MentS, TEQ, or RMET-R scores apart from intuition. Individuals that used intuition as their main strategy were found to perform worse across the entire experiment with a statistically significant negative correlation found between intuition and RMET-R scores, with a near statistically significant trend also found in the short and occluded conditions. Participants

who used intuition were also found to have lower TEQ scores in the short condition of the experiment but based off the results of the relationship between intuition and TEQ in other conditions this is most likely due to noise in the data.

The results of our study also showed that participants typically use multiple strategies. The participants of our experiment were found to use, on average, three strategies. Our inclusion of stimuli presentation time constraints was also found to significantly affect the number of strategies participants used, with participants using significantly less strategies in the occluded condition. Individuals that reported using more strategies were also found to have higher RMET-R scores, with a statistically significant correlation found between the number of strategies used and RMET-R scores in the short condition of the experiment and across the entire dataset. In the same manner individuals that used more strategies were found to score higher on the MentS as well, with a significant positive correlation found between the number of strategies used and MentS scores in the long condition of the experiment and across the entire dataset

The results of this study raise a number of important questions and implications regarding the use of the RMET-R, MentS, and TEQ. The following sections will go over the implications of this research for these measures, other ToM assessments, ToM researchers, and implications for the problem of compensation with ToM measures.

Implications for the RMET-R

Implications of our experimental manipulations

The findings of this study indicate that limiting the available answer time of the RMET-R offers a way to increase both the test's difficulty and its ecological validity. Although this is not what we expected this might prove quite useful as although the RMET-R is an exceedingly popular ToM assessment the test is easier than real world social situations,

to the point that it was been criticised as useless for assessing higher functioning individuals (Canty, 2016; Black, 2019). As the results of our short condition show, adding a five second answer time limit significantly increases the difficulty of the assessment, and makes the test far more akin to real-world affective ToM functioning where split-second decisions have to be repeatedly made in everyday conversation. This additional constraint is easy to add to the RMET-R and requires no additional resources, and as our results show does not affect the validity or reliability of the test. Due to this, this addition to the RMET-R is of great value to future researchers as it offers an easy way to improve two commonly noted issues with the RMET-R with no real drawback.

The results of this study also surprisingly indicate that adding stimulus presentation time limits does not significantly affect the difficulty of the RMET-R. This suggests that neurotypical individuals can work out the answer to the questions of the RMET-R with only a very brief glance. This has a number of implications for the RMET-R and other affective ToM measures, which will be discussed later in this chapter. Due to this lack of effect on the RMET-R the argument can easily be made that limiting the time that the RMET-R's pictures are visible can improve the RMET-R's ecological validity even further without significantly affecting the test's difficulty, validity or reliability. In the same manner these results also suggest that the base RMET-R has very poor ecological validity given that neurotypical individuals can complete the assessment after a half second glance with only a minor non statistically significant drop in results. Due to this, although adding tight stimuli visibility restrictions to the RMET-R does not significantly improve the test's difficulty, validity or reliability, it does allow one to improve the ecological validity of the assessment. This does require more effort to implement than adding increased answer time limits. In addition, judging by the results of our manipulation check variables the use of this additional constraint also increases the mental effort of the RMET-R, which forces participants to pay much closer

attention to the assessment and makes the assessment more mentally arousing. This increases the ecological validity of the assessment even further as it should theoretically increase the mental arousal and the engagement of participants. These are both issues that have been criticised with existing assessments, a number of assessments having been noted to have poor abilities in this regard (Bzdok et al., 2012).

Contrary to as we hypothesised neither of our time constraints increased the validity or reliability of the assessment. Due to our prior confound analyses it is extremely unlikely that this result is due to the presence of any form of suppressor variable and given our sizable sample it is unlikely there are any issues with the power of this study. However, this unexpected result may be due to our neurotypical sample, as we theorised that these constraints would improve the validity and reliability of the assessment by directly disrupting compensating individuals' abilities to use their compensatory strategies and processes. Which should theoretically cause them to score in a more consistent manner that more accurately reflects their actual ToM abilities, causing the construct validity and the internal consistency of the RMET-R to improve. However if our sample does not contain enough individuals that use compensation, there will naturally not be a significant effect. Unfortunately as we lacked access to a clinical sample of individuals which were likely to use compensation we were forced to instead use a large non-clinical sample to test our hypotheses, and hope that we managed to have a number of lower functioning individuals in our sample. Due to this it is quite possible that our lack of hypothesised results is simply due to a lack of compensating individuals in our sample. A reproduction of our study with a clinical sample of individuals that likely use compensation, such as a sample of older high functioning autistic individuals would be of great value and may show results more in line with what we originally hypothesised.

Implications for the psychometrics of the RMET-R

The results of our study also demonstrated that although they should be theoretically related, there was no relationship between RMET-R or MentS scores. Unexpectedly, no relationship existed between RMET-R scores and the MentS Other subcategory, which is effectively a self-answer questionnaire of ToM abilities. Due to the other psychometric issues found with the MentS in this study which are discussed later in this chapter and the results of previous research which supports the supposed validity of the RMET-R it is more likely that this lack of a relationship is due to an issue with the MentS.

Our results also indicate that the RMET-R has a questionably low level of internal consistency, with all three of our variations having a Cronbach's alpha of between 0.6-0.7. This issue is by no means isolated to our study, with the RMET-R having been found to have a wide range of internal consistency in prior research (Chen et al., 2017; Baron-Cohen et al., 2015). Due to this it is unlikely that this low level of internal consistency is due to our experimental constraints, which suggests that this is an ongoing issue with the RMET-R that researchers should be aware of.

Implications for how the RMET-R is performed

As expected, our qualitative assessment provided us with a solid investigation into how individuals performed the RMET-R. We found eight separate strategies that were used by individuals, one of which is more commonly used than the others. This assessment provides the first investigation into how this test is actually performed. Our investigation implied that nearly half of all individuals rely on the use of rules of thumb to quickly answer the assessment. Excepting intuition we found that the main strategy participants used did not significantly affect how well they did on the test. Individuals that used intuition or lacked a strategy did slightly worse on the overall test. We also found that most individuals used multiple strategies to complete the RMET-R, with participants reporting switching between or using in conjunction an average of three strategies to complete the assessment. Our results

also indicated that individuals that used more strategies typically did better and that individuals that relied on intuition did slightly worse. These findings, although they do not produce any ground-breaking revelations do provide a solid base into how the RMET-R is conducted for future researchers.

Implications for other measures

Mentalization Scale (MentS)

The results of our study indicate that the MentS has a number of psychometric issues. For instance, the Self subcategory of the MentS was found to be negatively correlated with the other two subcategories of the test, and the Motivation subcategory was found to have two questions which were negatively correlated with the rest of the assessment. This is a significantly different pattern of results to what was reported by Dimitrijević et al. (2018) and implies that the MentS has poor internal consistency and doesn't fit its supposed three factor structure. We also found that although the MentS was related to the TEQ it was not related to the RMET-R in the manner that it theoretically should, with none of the MentS subtests, including its Other subscales found to have a significant relationship with the RMET-R. These results imply that the MentS has poor level of validity and reliability, and as such the assessment requires some level of revision.

Toronto Empathy Questionnaire (TEQ)

The results of our study suggest that the TEQ has good psychometric properties, with the assessment found to have a good level of internal consistency and concurrent validity. However, although the TEQ was significantly related to the RMET-R in the manner that it should theoretically be the effect size was smaller than we expected, and the effect size that was reported by Spreng et al. (2009). This may suggest that the TEQ has a smaller relationship with the RMET-R than other empathy measures as nonsignificant results have

been reported before in other individuals' research (Woody, 2015). One other possibility is that the relationship between empathy and ToM may be significantly stronger in very high functioning and low functioning individuals, though this is beyond the scope of this research.

Marlowe-Crowne Social Desirability Scale Short-Form (MCSDS-SF)

Although it wasn't used as part of one of our analysis and we didn't assess the psychometric properties of the test we found the MCSDS-SF very useful for its designed purpose. The test successfully pointed out a number of individuals whose results suggested that they were faking good. Additionally, this short version of the test was easy to use and fit perfectly into our online experiment, exactly as it was designed to be used. The only issue we had with the assessment was that the MCSDS-SF lacked the norms of the MCSDS that are used to assess whether a participant was faking good. As we lacked this information, we selected to remove any participant that scored over 2SD on the MCSDS, which may have been too lenient, but it seemed to work for this study. This issue is something that should be examined in future research, as without the presence of any norms for this assessment it would be very difficult to standardise the use of this assessment.

Implications for other ToM assessments

Based on the results of our experiment it is possible that the effect of our experimental answer and stimulus presentation time constraints could easily generalise to other Tom assessments. Potentially offering an easy way to increase the difficulty and ecological validity of many other ToM assessments without unduly effecting their validity or reliability. Their validity and reliability could also be improved as well, depending on if our experimental manipulations affect compensating individuals in the manner that we theorised or not. Consequentially, investigating whether this effect holds up in other ToM assessments could be of great value.

Our results suggest that our experimental constraints are likely to be beneficial for affective ToM assessments. Since our results indicate that neurotypical individuals can respond to affective ToM tasks extremely fast, this suggests that almost all affective ToM assessments have very poor ecological validity. Those that rely on the use of static stimuli such as the new child reading mind in eyes test (van der Meulen et al., 2017) and yoni task (Shamay-Tsoory et al., 2006) having especially poor ecological validity. Although it requires further assessment our experimental time constraints should theoretically have the same effect on cognitive assessments as they do on affective assessments. Given the more complicated nature of cognitive ToM tasks it is possible that answer and stimuli time constraints might have a more pronounced effect. As such, future research into the effect of our experimental constraints on cognitive assessment may prove to be highly valuable and help to further illuminate the potential value of our time constraints.

With regard to video-based assessments the addition of increased answer time constraints would likely increase the assessments difficulty and ecological validity in the same manner of other assessments. However, given the active nature of the stimuli of video-based ToM assessments such as the Movie for the Assessment of Social Cognition (Dziobek et al., 2006) limiting the presentation time of the stimuli is not a viable tactic. Employing this form of constraint would by necessity require the recreation of the entire assessments' stimuli and would come at a great cost to time and effort. Our finding that neurotypical individuals can answer the RMET-R after only viewing its stimuli for a half second without any significant affect to the tests scores does suggest that participants can fairly easily answer the simple scenarios that are often used in video-based assessments. Due to this it is important for video-based assessments to use complex, and highly realistic scenarios that can appropriately mimic real-world social functioning.

Implications for the assessment of compensation

Contrary to as we hypothesised our study found no evidence that our experimental time constraints could either deny the use of compensatory strategies or be used to assess for evidence of compensation. As our experimental constraints had no significant effect on the validity of the RMET-R there was no point in conducting a simple slopes analysis to further explore how our time constraints moderated the relationship between the RMET-R and our predictor and outcome variables. However, as noted in the RMET-R section of this discussion this lack of a hypothesised effect may be due to a lack of compensating individuals in our sample. As such, our current results are inconclusive in regard to whether or not decreasing the answer and stimulus presentation time of ToM assessments can deny the use of compensatory strategies or allow one to search for compensation within ToM assessments. Due to this further research with a clinical sample of compensating individuals is necessary to investigate if our theory holds merit.

Implications for ToM researchers

The results of our study have a number of implications for ToM researchers. One major take from our study is the importance of considering answer time limits when developing or choosing assessments to use. The inclusion of an answer time allows for an easy way to manipulate the difficulty of the test and increases the ecologically validity of the assessment. Another implication of our study is that affective ToM abilities can be conducted extremely fast, as evidenced by the lack of any significant effect of our half second stimuli presentation constraint on the RMET-R's difficulty. This suggests that future affective ToM assessments should be developed to be faster and more difficult than they currently are, with even the currently available audio and video stimuli based affective assessments allowing significantly more time than our findings suggest neurotypical individual actually require (Golan et al., 2005; Dziobek et al., 2006).

The findings of our qualitative analysis also provide a spread of information on how affective tasks such as the RMET-R are performed. Our findings indicate that when assessing others' immediate mental states the majority of individuals will fall back on the use of rules of thumb such as eye direction to quickly make initial judgements. However, the results of our study also indicated that the majority of individuals will commonly switch between a number of strategies depending on the difficulty of the judgement call, with the participants of this study found to routinely switch between the use of rules of thumb, their intuition, mental imagery, their previous experiences, and their general knowledge, or use multiple strategies in conjunction to discern the mental states of other individuals. Our study also suggested that individuals who chose to rely on their intuition or didn't attempt to switch between strategies were significantly worse at working out the mental states of other individuals. Interestingly, we found that our half a second stimuli presentation time constraint had an effect on some individuals' abilities to use rules of thumb, with significantly less individuals found to use this as their main strategy in the occluded condition of our experiment. This suggests that some individuals can use rules of thumb faster than others. These findings do not provide any great revelations but do provide an interesting look into how individuals actually perform affective ToM tasks and provide a useful basis of information for future researchers.

The success of our increased answer time constraints in increasing the difficulty of the assessment also implies that it is possible to create a multi timed assessment with multiple difficulty levels. This kind of assessment could be of great value as it could allow for a much more accurate assessment of an individual's true ToM abilities. No ToM assessments currently exist for adults that allow for the of assessment of multiple levels of ToM ability. One last implication of our study for other ToM researchers is the high value of Prolific for

gathering non-clinical samples of individuals. Our study easily gathered 381 participants within two days.

Strengths and general limitations of study

As this study was designed to investigate a theory that had never been previously assessed this study was highly experimental in nature and ended up having a number of significant strengths and limitations. In respect to the strengths of the study, apart from the MentS not having a relationship with the RMET-R that we theorised it would our design worked flawlessly. All of the segments of our programmed experiment worked well together, and the only issue noted by our participants was the difficulty of our experimental manipulations. Our experiment was noted to be highly interesting and engrossing by multiple participants, and despite its size didn't take long to complete. Our use of Prolific also allowed us to gather a sample size more than double the 184 participants that our priori power analysis suggested we needed to have a power of .90 for our theorised effect size of 0.07 which resulted in our study having plenty of power. This large sample size was gathered in part to try to make up for our lack of a clinical sample, as our experimental constraints would only theoretically increase the validity and reliability of the RMET-R if we had a number of compensating individuals in our study. The attention checks and the MCSDS-SF we included also functioned well and allowed us to identify and exclude a number of biased data sets.

As previously noted one of the greatest limitations of our current study was its convenience sample of neurotypical individuals, which was not particularly well suited to assess our main theory. Unfortunately as we lacked any access to a clinical sample the best we could do was use a large online sample and hope that our sample managed to include some individuals that used compensatory techniques. Against our expectations we also discovered that the MentS was not related to the RMET-R in the manner that we theorised. Due to this we could only use the TEQ to assess if our experimental constraints increased the

validity of the RMET-R which limited our ability to assess our experimental constraints. One other limitation was our high reliance on self-report measures such as the TEQ and MentS which can be effected by social desirability bias, for this study we added the MCSDS-SF to help combat this issue, though it is possible that we may have been a bit too liberal in our exclusionary criteria with this test. As we conducted an online study we also had less control over our study conditions, though due to the design of our study and the fact that recruited from a wide variety of individuals across four countries this shouldn't have been any other unexpected variables that might have effected the results of our study.

Future directions

The findings of this study, and our above discussion suggest that a variety of future research is required to clarity if decreasing the answer and stimuli presentation time of ToM assessments such as the RMET-R can deny compensating individuals the use of their compensatory strategies and improve the difficulty, validity and reliability of the assessment. Future research should replicate this study with a clinical sample of individuals that are likely to use compensation, which would enable a much clearer picture of the viability of our theory. To assess if our results hold up with other assessments future research should also attempt to replicate our study with other ToM assessments. With our experimental constraints potentially offering a way to improve the difficulty and ecological validity of many assessments without significantly effecting their psychometric properties.

Concluding remarks

This current study aimed to assess if decreasing the answer and stimuli presentation time of the RMET-R might offer a way to stop compensating individuals from using compensatory strategies to pass the assessment, which theoretically would lead to an increase in the assessment's difficulty, validity, and reliability. In contrast to as we predicted we found

that decreasing the available answer time of the RMET-R significantly increased the assessments difficulty but had no significant effect on its validity or reliability. And that decreasing the time that the RMET-R stimuli was presented had no significant effect on the assessment's difficulty, validity or reliability. This failure to find a significant effect between our experimental constraints and the validity and reliability of the RMET-R might be due to a lack of compensating individuals in our sample. As our theorised increase in the psychometric properties of the RMET-R came directly from forcing compensating individuals to answer in a more consistent and valid manner. As we unfortunately lacked access to a clinical sample we attempted to recruit a large online sample with the hope that our sample would end up including a number of compensating individuals. However due to this we cannot conclusively say whether the lack of a significant effect of our experimental manipulations on the validity and reliability of the RMET-R is because our theory is incorrect or because we just didn't have enough compensating individuals to have any kind of effect. Due to this further research with a clinical sample of compensating individuals is required to assess if our theory is correct.

Although this pattern of results was not what we expected our results suggest that our answer and stimulus presentation time constraints may actually prove useful for increasing the ecological validity of ToM assessments as they demonstrate no evidence of significantly effecting the psychometric properties of the RMET-R. Although further research is required to investigate if this effect is isolated to the RMET-R or holds up with other assessments, our two experimental constraints may prove highly useful for this alternative purpose. Although of less significance our study also presents a first look into how the RMET-R is actually performed, with our study creating a base of information for any other researchers that could find this information useful.

References

Abell, F., Happe, 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

Adams Jr, R. B., Rule, N. O., Franklin Jr, 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. (2001). The neurobiology of social cognition. *Current Opinion in Neurobiology, 11*(2), 231-239. https://doi.org/10.1016/S0959-4388(00)00202-6

Ahmed, F. S., & Miller, L. S. (2011). Executive function mechanisms of theory of mind. *Journal of Autism and Developmental Disorders, 41*(5), 667-678. https://doi.org/10.1007/s10803-010-1087-7

Ahmed, F. S., & Miller, L. S. (2013). Relationship between theory of mind and functional independence is mediated by executive function. *Psychology and Aging, 28*(2), 293-303. https://doi.org/10.1037/a0031365

Apperly, I. (2013). Can theory of mind grow up? Mindreading in adults, and its implications for the development and neuroscience of mindreading. In S. Baron-Cohen., M. Lombardo & H. Tager-Flusberg (Eds.), *Understanding other minds: Perspectives from developmental social neuroscience*, (3rd ed., pp. 72-92). Oxford University Press.

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

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

Bachmann, T., & Francis, G. (2013). *Visual masking: Studying perception, attention, and consciousness*: Academic Press.

Baddeley, A. (2003). Working memory: looking back and looking forward. *Nature Reviews Neuroscience, 4*(10), 829-839. https://doi.org/10.1038/nrn1201

Balaban, N., Friedmann, N., & Ariel, M. (2016a). The effect of theory of mind impairment on language: Referring after right-hemisphere damage. *Aphasiology, 30*(12), 1424-1460. https://doi.org/10.1080/02687038.2015.1137274

Balaban, N., Friedmann, N., & Ziv, M. (2016b). Theory of mind impairment after right-hemisphere damage. *Aphasiology, 30*(12), 1399-1423. https://doi.org/10.1080/02687038.2015.1137275

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. (1991a). Precursors to a theory of mind: Understanding attention in others. *Natural theories of mind: Evolution, development and simulation of everyday mindreading, 1*, 233-251.

Baron-Cohen, S. (1991b). The theory of mind deficit in autism: How specific is it? *British Journal of Developmental Psychology, 9*(2), 301-314. https://doi.org/10.1111/j.2044-835X.1991.tb00879.x

Baron-Cohen, S. (2000). Theory of mind and autism: A fifteen year review. *Understanding other minds: Perspectives from developmental cognitive neuroscience, 2*, 3-20.

Baron-Cohen, S., Bowen, D. C., Holt, R. J., Allison, C., Auyeung, B., Lombardo, M. V., Smith, P., & Lai, M.-C. (2015). The “reading the mind in the eyes” test: complete absence of typical sex difference in~ 400 men and women with autism. *PloS One, 10*(8), e0136521. https://doi.org/10.1371/journal.pone.0136521

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. (2001A). The “Reading the Mind in the Eyes” Test revised version: a study with normal adults, and adults with Asperger syndrome or high-functioning autism. *The Journal of Child Psychology and Psychiatry and Allied Disciplines, 42*(2), 241-251. https://doi.org/10.1017/S0021963001006643

Baron-Cohen, S., Wheelwright, S., Spong, A., Scahill, V., & Lawson, J. (2001B). Are intuitive physics and intuitive psychology independent? A test with children with Asperger Syndrome. *Journal of Developmental and Learning Disorders, 5*(1), 47-78.

Baron-Cohen, S., Willey, C. W., Grandin, T., & Jolliffe, T. (2000). Is high-functioning autism/Asperger's syndrome necessarily a disability. *Development and Psychopathology, 12*(3), 489-500. https://doi.org/10.1017/S0954579400003126

Beaumont, R. B., & Sofronoff, K. (2008). A new computerised advanced theory of mind measure for children with Asperger syndrome: The ATOMIC. *Journal of Autism and Developmental Disorders, 38*(2), 249-260. https://doi.org/10.1007/s10803-007-0384-2

Black, J. E. (2019). An IRT analysis of the Reading the Mind in the Eyes test. *Journal of Personality Assessment, 101*(4), 425-433. https://doi.org/10.1080/00223891.2018.1447946

Blackburn, P., Braüner, T., & Polyanskaya, I. (2015). *Second-order false-beliefs, language and logic.* [Paper presentation] Conference on Computing Natural Reasoning 2015, Bloomington, United States Of America.

Bodner, K. E., Engelhardt, C. R., Minshew, N. J., & Williams, D. L. (2015). Making inferences: Comprehension of physical causality, intentionality, and emotions in discourse by high-functioning older children, adolescents, and adults with autism. *Journal of Autism and Developmental Disorders, 45*(9), 2721-2733. https://doi.org/10.1007/s10803-015-2436-3

Bohl, V., & van den Bos, W. (2012). Toward an integrative account of social cognition: marrying theory of mind and interactionism to study the interplay of Type 1 and Type 2 processes. *Frontiers in Human Neuroscience, 6*, 274. https://doi.org/10.3389/fnhum.2012.00274

Bora, E., Bartholomeusz, C., & Pantelis, C. (2016). Meta-analysis of Theory of Mind (ToM) impairment in bipolar disorder. *Psychological Medicine, 46*(2), 253-264. https://doi.org/10.1017/S0033291715001993

Bora, E., & Berk, M. (2016). Theory of mind in major depressive disorder: A meta-analysis. *Journal of affective disorders, 191,* 49-55. https://doi.org/10.1016/j.jad.2015.11.023

Bora, E., Yucel, M., & Pantelis, C. (2009). Theory of mind impairment in schizophrenia: meta-analysis. *Schizophrenia Research, 109*(1-3), 1-9. https://doi.org/10.1016/j.schres.2008.12.020

Bosco, F. M., Capozzi, F., Colle, L., Marostica, P., & Tirassa, M. (2013). Theory of mind deficit in subjects with alcohol use disorder: an analysis of mindreading processes. *Alcohol and Alcoholism, 49*(3), 299-307. https://doi.org/10.1093/alcalc/agt148

Bosco, F. M., Colle, L., De Fazio, S., 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

Bosco, F. M., Gabbatore, I., & Tirassa, M. (2014). A broad assessment of theory of mind in adolescence: the complexity of mindreading. *Consciousness and Cognition, 24*, 84-97. https://doi.org/10.1016/j.concog.2014.01.003

Bosco, F. M., Gabbatore, I., Tirassa, M., & Testa, S. (2016). Psychometric properties of the Theory of Mind Assessment Scale in a sample of adolescents and adults. *Frontiers in Psychology, 7*, 566. https://doi.org/10.3389/fpsyg.2016.00566

Bossaerts, P., Suzuki, S., & O'Doherty, J. P. (2019). Perception of intentionality in investor attitudes towards financial risks. *Journal of Behavioral and Experimental Finance, 23*, 189-197. https://doi.org/10.1016/j.jbef.2017.12.011

Brent, E., Rios, P., Happé, F., & Charman, T. (2004). Performance of children with autism spectrum disorder on advanced theory of mind tasks. *Autism, 8*(3), 283-299. https://doi.org/10.1177/1362361304045217

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

Brüne, M. (2005). "Theory of mind" in schizophrenia: a review of the literature. *Schizophrenia Bulletin, 31*(1), 21-42. https://doi.org/10.1093/schbul/sbi002

Brüne, M., & Brüne-Cohrs, U. (2006). Theory of mind—evolution, ontogeny, brain mechanisms and psychopathology. *Neuroscience & Biobehavioral Reviews, 30*(4), 437-455. https://doi.org/10.1016/j.neubiorev.2005.08.001

Bugnyar, T., Reber, S. A., & Buckner, C. (2016). Ravens attribute visual access to unseen competitors. *Nature Communications, 7*(1), 1-6. https://doi.org/10.1038/ncomms10506

Byom, L. J., & Mutlu, B. (2013). Theory of mind: Mechanisms, methods, and new directions. *Frontiers in Human Neuroscience, 7*, 413. https://doi.org/10.3389/fnhum.2013.00413

Bzdok, D., Schilbach, L., Vogeley, K., Schneider, K., Laird, A. R., Langner, R., & Eickhoff, S. B. (2012). Parsing the neural correlates of moral cognition: ALE meta-analysis on morality, theory of mind, and empathy. *Brain Structure and Function, 217*(4), 783-796. https://doi.org/10.1007/s00429-012-0380-y

Cage, E., Di Monaco, J., & Newell, V. (2018). Experiences of autism acceptance and mental health in autistic adults. *Journal of Autism and Developmental Disorders, 48*(2), 473-484. https://doi.org/10.1007/s10803-017-3342-7

Canty, A. L. (2016). *The Grinch Who Stole Thoughts: A Virtual Reality Study of Theory of Mind in Early Psychosis and Chronic Schizophrenia.* [Doctoral dissertation, Griffith University] Griffith Research Online https://research-repository.griffith.edu.au/bitstream/handle/10072/368165/Canty_2016_01Thesis.pdf?sequence=1

Canty, A. L., Neumann, D. L., Fleming, J., & Shum, D. H. (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

Carruthers, P. (2016). Two systems for mindreading? *Review of Philosophy and Psychology, 7*(1), 141-162. https://doi.org/10.1007/s13164-015-0259-y

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

Chandler, M. J. (1973). Egocentrism and antisocial behavior: The assessment and training of social perspective-taking skills. *Developmental Psychology, 9*(3), 326-332. https://doi.org/10.1037/h0034974

Chen, K.-W., Lee, S.-C., Chiang, H.-Y., Syu, Y.-C., Yu, X.-X., & Hsieh, C.-L. (2017). Psychometric properties of three measures assessing advanced theory of mind: Evidence from people with schizophrenia. *Psychiatry Research, 257*, 490-496. https://doi.org/10.1016/j.psychres.2017.08.026

Christensen, W., & Michael, J. (2016). From two systems to a multi-systems architecture for mindreading. *New Ideas in Psychology, 40*, 48-64. https://doi.org/10.1016/j.newideapsych.2015.01.003

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

Crane, L., Goddard, L., & Pring, L. (2013). Autobiographical memory in adults with autism spectrum disorder: The role of depressed mood, rumination, working memory and theory of mind. *Autism, 17*(2), 205-219. https://doi.org/10.1177/1362361311418690

Crowne, D. P., & Marlowe, D. (1960). A new scale of social desirability independent of psychopathology. *Journal of Consulting Psychology, 24*(4), 349–354. https://doi.org/10.1037/h0047358

Dal Monte, O., Schintu, S., Pardini, M., Berti, A., Wassermann, E. M., Grafman, J., & Krueger, F. (2014). The left inferior frontal gyrus is crucial for reading the mind in the eyes: brain lesion evidence. *Cortex, 58*, 9-17. https://doi.org/10.1016/j.cortex.2014.05.002

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

Dimitrijević, A., Hanak, N., Altaras Dimitrijević, A., & Jolić Marjanović, Z. (2018). The Mentalization Scale (MentS): A self-report measure for the assessment of mentalizing capacity. *Journal of Personality Assessment, 100*(3), 268-280. https://doi.org/10.1080/00223891.2017.1310730

Dimopoulou, T., Tarazi, F. I., & Tsapakis, E. M. (2017). Clinical and therapeutic role of mentalization in schizophrenia—a review. *CNS Spectrums, 22*(6), 450-462. https://doi.org/10.1017/S1092852916000687

Dodell-Feder, D., Ressler, K. J., & Germine, L. T. (2020). Social cognition or social class and culture? On the interpretation of differences in social cognitive performance. *Psychological Medicine, 50*(1), 133-145. https://doi.org/10.1017/S003329171800404X

Domes, G., Heinrichs, M., Michel, A., Berger, C., & Herpertz, S. C. (2007). Oxytocin improves “mind-reading” in humans. *Biological Psychiatry, 61*(6), 731-733. https://doi.org/10.1016/j.biopsych.2006.07.015

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

Dunbar, R. I. (2003). The social brain: mind, language, and society in evolutionary perspective. *Annual review of Anthropology, 32*(1), 163-181. https://doi.org/10.1146/annurev.anthro.32.061002.093158

Durrleman, S., & Franck, J. (2015). Exploring links between language and cognition in autism spectrum disorders: Complement sentences, false belief, and executive functioning. *Journal of Communication Disorders, 54*, 15-31. https://doi.org/10.1016/j.jcomdis.2014.12.001

Dvash, J., & Shamay-Tsoory, S. G. (2014). Theory of mind and empathy as multidimensional constructs: Neurological foundations. *Topics in Language Disorders, 34*(4), 282-295. https://doi.org/10.1097/TLD.0000000000000040

Dziobek, I., Fleck, S., Kalbe, E., Rogers, K., Hassenstab, J., Brand, M., Kessler, J., Woike, J., Wolf, O., & 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

Eddy, C., Sira Mahalingappa, S., & Rickards, H. (2012). Is Huntington's disease associated with deficits in theory of mind? *Acta Neurologica Scandinavica, 126*(6), 376-383. https://doi.org/10.1111/j.1600-0404.2012.01659.x

Enright, R. D., & Lapsley, D. K. (1980). Social role-taking: A review of the constructs, measures, and measurement properties. *Review of Educational Research, 50*(4), 647-674. https://doi.org/10.3102/00346543050004647

Eyuboglu, M., Baykara, B., & Eyuboglu, D. (2018). Broad autism phenotype: theory of mind and empathy skills in unaffected siblings of children with autism spectrum disorder. *Psychiatry and Clinical Psychopharmacology, 28*(1), 36-42. https://doi.org/10.1080/24750573.2017.1379714

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods, 39*, 175-191. https://doi.org/10.3758/BF03193146

Feffer, M. H., & Gourevitch, V. (1960). Cognitive aspects of role-taking in children 1. *Journal of Personality, 28*(4), 383-396. https://doi.org/10.1111/j.1467-6494.1960.tb01627.x

Ferguson, F. J., & Austin, E. J. (2010). Associations of trait and ability emotional intelligence with performance on Theory of Mind tasks in an adult sample. *Personality and Individual Differences*, *49*(5), 414-418. https://doi.org/10.1016/j.paid.2010.04.009

Fiske, S. T., & Taylor, S. E. (2013). *Social cognition: From brains to culture*. Sage.

Fonagy, P., & Allison, E. (2014). The role of mentalizing and epistemic trust in the therapeutic relationship. *Psychotherapy, 51*(3), 372-380. https://doi.org/10.1037/a0036505

Fossati, A., Somma, A., Krueger, R. F., Markon, K. E., & Borroni, S. (2017). On the relationships between DSM-5 dysfunctional personality traits and social cognition deficits: A study in a sample of consecutively admitted Italian psychotherapy patients. *Clinical Psychology & Psychotherapy, 24*(6), 1421-1434. https://doi.org/10.1002/cpp.2091

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

Gallese, V., & Goldman, A. (1998). Mirror neurons and the simulation theory of mind-reading. *Trends in Cognitive Sciences, 2*(12), 493-501. https://doi.org/10.1016/S1364-6613(98)01262-5

Gerrans, P. (2002). The theory of mind module in evolutionary psychology. *Biology and Philosophy, 17*(3), 305-321. https://doi.org/10.1023/A:1020183525825

Golan, O., Baron-Cohen, S., & Hill, J. (2006a). 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. (2006b). 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

Green, M. F., Horan, W. P., & Lee, J. (2015). Social cognition in schizophrenia. *Nature Reviews Neuroscience, 16*(10), 620-631. https://doi.org/10.1038/nrn4005

Green, M. F., Olivier, B., Crawley, J. N., Penn, D. L., & Silverstein, S. (2005). Social cognition in schizophrenia: recommendations from the measurement and treatment research to improve cognition in schizophrenia new approaches conference. *Schizophrenia Bulletin, 31*(4), 882-887. https://doi.org/10.1093/schbul/sbi049

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

Hair, J. F., Anderson, R. E., Babin, B. J., & Black, W. C. (2010). *Multivariate data analysis: A global perspective* (7th ed.). Pearson.

Hamilton, J., Radlak, B., Morris, P. G., & Phillips, L. H. (2017). Theory of mind and executive functioning following stroke. *Archives of Clinical Neuropsychology, 32*(5), 507-518. https://doi.org/10.1093/arclin/acx035

Happé, F., Cook, J. L., & Bird, G. (2017). The structure of social cognition: In (ter) dependence of sociocognitive processes. *Annual Review of Psychology, 68*, 243-267. https://doi.org/10.1146/annurev-psych-010416-044046

Happé, F., & Frith, U. (2014). Annual research review: Towards a developmental neuroscience of atypical social cognition. *Journal of Child Psychology and Psychiatry, 55*(6), 553-577. https://doi.org/10.1111/jcpp.12162

Happé, F. G. (1994). An advanced test of theory of mind: Understanding of story characters' thoughts and feelings by able autistic, mentally handicapped, and normal children and

adults. *Journal of Autism and Developmental Disorders, 24*(2), 129-154. https://doi.org/10.1007/BF02172093

Hardy, M. A. (1993). *Regression with dummy variables* (Sage university paper series on quantitative applications in the social sciences, series no. 07-093). Sage.

Hayward, E. O., & Homer, B. D. (2017). Reliability and validity of advanced theory-of-mind measures in middle childhood and adolescence. *British Journal of Developmental Psychology, 35*(3), 454-462. https://doi.org/10.1111/bjdp.12186

Heavey, L., Phillips, W., Baron-Cohen, S., & Rutter, M. (2000). The Awkward Moments Test: A naturalistic measure of social understanding in autism. *Journal of Autism and Developmental Disorders, 30*(3), 225-236. https://doi.org/10.1023/A:1005544518785

Hollebrandse, B., van Hout, A., & Hendriks, P. (2014). Children's first and second-order false-belief reasoning in a verbal and a low-verbal task. *Synthese, 191*(3), 321-333. https://doi.org/10.1007/s11229-012-0169-9

Hull, L., Mandy, W., Lai, M.-C., Baron-Cohen, S., Allison, C., Smith, P., & Petrides, K. (2019). Development and validation of the camouflaging autistic traits questionnaire (CAT-Q). *Journal of Autism and Developmental Disorders, 49*(3), 819-833. https://doi.org/10.1007/s10803-018-3792-6

Hull, L., Mandy, W., & Petrides, K. (2017a). Behavioural and cognitive sex/gender differences in autism spectrum condition and typically developing males and females. *Autism, 21*(6), 706-727. https://doi.org/10.1177/1362361316669087

Hull, L., Petrides, K., Allison, C., Smith, P., Baron-Cohen, S., Lai, M.-C., & Mandy, W. (2017b). "Putting on my best normal": social camouflaging in adults with autism spectrum conditions. *Journal of Autism and Developmental Disorders, 47*(8), 2519-2534. https://doi.org/10.1007/s10803-017-3166-5

Hutchins, T. L., Prelock, P. A., & Bonazinga, L. (2012). Psychometric evaluation of the Theory of Mind Inventory (ToMI): A study of typically developing children and children with autism spectrum disorder. *Journal of Autism and Developmental Disorders, 42*(3), 327-341. https://doi.org/10.1007/s10803-011-1244-7

Im-Bolter, N., Agostino, A., & Owens-Jaffray, K. (2016). Theory of mind in middle childhood and early adolescence: Different from before? *Journal of Experimental Child Psychology, 149*, 98-115. https://doi.org/10.1016/j.jecp.2015.12.006

Inoue, Y., Yamada, K., & Kanba, S. (2006). Deficit in theory of mind is a risk for relapse of major depression. *Journal of Affective Disorders, 95*(1-3), 125-127. https://doi.org/10.1016/j.jad.2006.04.018

Joffe, T. H. (1997). Social pressures have selected for an extended juvenile period in primates. *Journal of Human Evolution, 32*(6), 593-605. https://doi.org/10.1006/jhev.1997.0140

Johnson, M. H., Jones, E. J., & Gliga, T. (2015). Brain adaptation and alternative developmental trajectories. *Development and Psychopathology, 27*(2), 425-442. https://doi.org/10.1017/S0954579415000073

Jones, C. R., Simonoff, E., Baird, G., Pickles, A., Marsden, A. J., Tregay, J., Happé, F., & Charman, T. (2018). The association between theory of mind, executive function, and the symptoms of autism spectrum disorder. *Autism Research, 11*(1), 95-109. https://doi.org/10.1002/aur.1873

Kaiser, M. D., Delmolino, L., Tanaka, J. W., & Shiffrar, M. (2010a). Comparison of visual sensitivity to human and object motion in autism spectrum disorder. *Autism Research, 3*(4), 191-195. https://doi.org/10.1002/aur.137

Kaiser, M. D., Hudac, C. M., Shultz, S., Lee, S. M., Cheung, C., Berken, A. M., Deen, B., Pitskel, N. B., Sugrue, D. R., Voos, A. C., Saulnier, C. A. (2010b). Neural signatures of autism. *Proceedings of the National Academy of Sciences, 107*(49), 21223-21228. https://doi.org/10.1073/pnas.1010412107

Kaland, N., Callesen, K., Møller-Nielsen, A., Mortensen, E. L., & Smith, L. (2008). Performance of children and adolescents with Asperger syndrome or high-functioning autism on advanced theory of mind tasks. *Journal of Autism and Developmental Disorders, 38*(6), 1112-1123. https://doi.org/10.1007/s10803-007-0496-8

Kaland, N., Møller-Nielsen, A., Callesen, K., Mortensen, E. L., Gottlieb, D., & Smith, L. (2002). A new advanced test of theory of mind: evidence from children and adolescents with Asperger syndrome. *Journal of Child Psychology and Psychiatry, 43*(4), 517-528. https://doi.org/10.1111/1469-7610.00042

Kana, R. K., Maximo, J. O., Williams, D. L., Keller, T. A., Schipul, S. E., Cherkassky, V. L., Minshew, N. J., Just, M. A. (2015). Aberrant functioning of the theory-of-mind network in children and adolescents with autism. *Molecular Autism, 6*(1), 59. https://doi.org/10.1186/s13229-015-0052-x

Kanske, P., Böckler, A., & Singer, T. (2015). Models, mechanisms and moderators dissociating empathy and theory of mind. In M. Wöhr & S. Krach (Eds.), *Social Behavior from Rodents to Humans: Neural Foundations and Clinical Implications* (pp. 193-206): Springer.

Kantak, S. S., Stinear, J. W., Buch, E. R., & Cohen, L. G. (2012). Rewiring the brain: potential role of the premotor cortex in motor control, learning, and recovery of function following brain injury. *Neurorehabilitation and Neural Repair, 26*(3), 282-292. https://doi.org/10.1177/1545968311420845

Kawulich, B. B. (2004). Data analysis techniques in qualitative research. *Journal of Research in Education, 14*(1), 96-113.

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

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

Kipps, C. M., & Hodges, J. R. (2006). Theory of mind in frontotemporal dementia. *Social Neuroscience, 1*(3-4), 235-244. https://doi.org/10.1080/17470910600989847

Kirkland, R. A., Peterson, E., Baker, C. A., Miller, S., & Pulos, S. (2013). Meta-analysis Reveals Adult Female Superiority in" Reading the Mind in the Eyes Test". *North American Journal of Psychology*, *15*(1). 121-146.

Klein, A. M., Zwickel, J., Prinz, W., & Frith, U. (2009). Animated triangles: An eye tracking investigation. *The Quarterly Journal of Experimental Psychology, 62*(6), 1189-1197. https://doi.org/10.1080/17470210802384214

Klin, A. (2000). Attributing social meaning to ambiguous visual stimuli in higher-functioning autism and Asperger syndrome: The social attribution task. *The Journal of Child Psychology and Psychiatry and Allied Disciplines, 41*(7), 831-846. https://doi.org/10.1111/1469-7610.00671

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

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

Kuntoro, I. A., Peterson, C. C., & Slaughter, V. (2017). Culture, parenting, and children's theory of mind development in Indonesia. *Journal of Cross-Cultural Psychology, 48*(9), 1389-1409. https://doi.org/10.1177/0022022117725404

Laghi, F., Cotugno, A., Cecere, F., Sirolli, A., Palazzoni, D., & Bosco, F. M. (2014). An exploratory assessment of theory of mind and psychological impairment in patients with bulimia nervosa. *British Journal of Psychology, 105*(4), 509-523. https://doi.org/10.1111/bjop.12054

Lagravinese, G., Avanzino, L., Raffo De Ferrari, A., Marchese, R., Serrati, C., Mandich, P., Abbruzzese, G., Pelosin, E. (2017). Theory of mind is impaired in mild to moderate Huntington's disease independently from global cognitive functioning. *Frontiers in Psychology, 8*, 80. https://doi.org/10.3389/fpsyg.2017.00080

Lai, M.-C., Lombardo, M. V., Auyeung, B., Chakrabarti, B., & Baron-Cohen, S. (2015). Sex/gender differences and autism: setting the scene for future research. *Journal of the American Academy of Child & Adolescent Psychiatry, 54*(1), 11-24. https://doi.org/10.1016/j.jaac.2014.10.003

Lai, M.-C., Lombardo, M. V., Chakrabarti, B., Ruigrok, A. N., Bullmore, E. T., Suckling, J., Auyeung, B., Happé, F., Szatmari, P., & Baron-Cohen, S. (2019). Neural self-representation in autistic women and association with 'compensatory camouflaging'. *Autism, 23*(5), 1210-1223. https://doi.org/10.1177/1362361318807159

Lai, M.-C., Lombardo, M. V., Pasco, G., Ruigrok, A. N., Wheelwright, S. J., Sadek, S. A Chakrabarti, B., Baron-Cohen, S., & MRC AIMS Consortium. (2011). A behavioral

comparison of male and female adults with high functioning autism spectrum conditions. *PloS One, 6*(6), e20835. https://doi.org/10.1371/journal.pone.0020835

Lambert, C. E., Arbuckle, S. A., & Holden, R. R. (2016). The Marlowe–Crowne social desirability scale outperforms the BIDR impression management scale for identifying fakers. *Journal of Research in Personality, 61*, 80-86. https://doi.org/10.1016/j.jrp.2016.02.004

Lamm, C., Decety, J., & Singer, T. (2011). Meta-analytic evidence for common and distinct neural networks associated with directly experienced pain and empathy for pain. *Neuroimage, 54*(3), 2492-2502. https://doi.org/10.1016/j.neuroimage.2010.10.014

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

Lehnhardt, F.-G., Falter, C. M., Gawronski, A., Pfeiffer, K., Tepest, R., Franklin, J., & Vogeley, K. (2016). Sex-related cognitive profile in autism spectrum disorders diagnosed late in life: implications for the female autistic phenotype. *Journal of Autism and Developmental Disorders, 46*(1), 139-154. https://doi.org/10.1007/s10803-015-2558-7

Leudar, I., Costall, A., & Francis, D. (2004). Theory of mind: a critical assessment. *Theory & Psychology, 14*(5), 571-578. https://doi.org/10.1177/0959354304046173

Liu, D., Wellman, H. M., Tardif, T., & Sabbagh, M. A. (2008). Theory of mind development in Chinese children: a meta-analysis of false-belief understanding across cultures and languages. *Developmental Psychology, 44*(2), 523-531. https://doi.org/10.1037/0012-1649.44.2.523

Livingston, L. A., Carr, B., & Shah, P. (2019a). Recent advances and new directions in measuring theory of mind in autistic adults. *Journal of Autism and Developmental Disorders, 49*(4), 1738-1744. https://doi.org/10.1007/s10803-018-3823-3

Livingston, L. A., Colvert, E., Team, S. R. S., Bolton, P., & Happé, F. (2019b). Good social skills despite poor theory of mind: exploring compensation in autism spectrum disorder. *Journal of Child Psychology and Psychiatry, 60*(1), 102-110. https://doi.org/10.1111/jcpp.12886

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

Livingston, L. A., Shah, P., & Happé, F. (2019c). Compensatory strategies below the behavioural surface in autism: a qualitative study. *The Lancet Psychiatry, 6*(9), 766-777. https://doi.org/10.1016/S2215-0366(19)30224-X

Low, J., & Perner, J. (2012). Implicit and explicit theory of mind: State of the art. *British Journal of Developmental Psychology, 30*(1), 1-13. https://doi.org/10.1111/j.2044-835X.2011.02074.x

Maginnity, M. E., & Grace, R. C. (2014). Visual perspective taking by dogs (Canis familiaris) in a Guesser–Knower task: evidence for a canine theory of mind? *Animal Cognition, 17*(6), 1375-1392. https://doi.org/10.1007/s10071-014-0773-9

Maniaci, M. R., & Rogge, R. D. (2014). Caring about carelessness: Participant inattention and its effects on research. *Journal of Research in Personality, 48*, 61-83. https://doi.org/10.1016/j.jrp.2013.09.008

Martory, M.-D., Pegna, A. J., Sheybani, L., Métral, M., Pertusio, F. B., & Annoni, J.-M. (2015). Assessment of social cognition and theory of mind: initial validation of the

Geneva Social Cognition Scale. *European Neurology, 74*(5-6), 288-295. https://doi.org/10.1159/000442412

McAlister, A., & Peterson, C. (2007). A longitudinal study of child siblings and theory of mind development. *Cognitive Development, 22*(2), 258-270. https://doi.org/10.1016/j.cogdev.2006.10.009

McDonald, S., Flanagan, S., Rollins, J., & Kinch, J. (2003). TASIT: A new clinical tool for assessing social perception after traumatic brain injury. *The Journal of Head Trauma Rehabilitation, 18*(3), 219-238. https://doi.org/10.1097/00001199-200305000-00001

McPartland, J. C. (2019). Autism's existential crisis: a reflection on Livingston et al.(2018). *Journal of Child Psychology and Psychiatry, 60*(1), 111-113. https://doi.org/10.1111/jcpp.12989

Mehta, U. M., Thirthalli, J., Kumar, C. N., Mahadevaiah, M., Rao, K., Subbakrishna, D. K., Gangadhar, B.N., & Keshavan, M. S. (2011). Validation of Social Cognition Rating Tools in Indian Setting (SOCRATIS): A new test-battery to assess social cognition. *Asian Journal of Psychiatry, 4*(3), 203-209. https://doi.org/10.1016/j.ajp.2011.05.014

Mena, B., José, M., Alarcón, R., Arnau Gras, J., Bono Cabré, R., & Bendayan, R. (2017). Non-normal data: Is ANOVA still a valid option?. *Psicothema, 29*(4), 552-557. https://doi.org/10.7334/psicothema2016.383

Milders, M., Fuchs, S., & Crawford, J. R. (2003). Neuropsychological impairments and changes in emotional and social behaviour following severe traumatic brain injury. *Journal of Clinical and Experimental Neuropsychology, 25*(2), 157-172. https://doi.org/10.1076/jcen.25.2.157.13642

Milligan, K., Astington, J. W., & Dack, L. A. (2007). Language and theory of mind: Meta-analysis of the relation between language ability and false-belief understanding. *Child Development, 78*(2), 622-646. https://doi.org/10.1111/j.1467-8624.2007.01018.x

Mintah, K., & Parlow, S. E. (2018). Are you flirting with me? Autistic traits, theory of mind, and inappropriate courtship. *Personality and Individual Differences, 128,* 100-106. https://doi.org/10.1016/j.paid.2018.02.028

Miu, A. C., Pană, S. E., & Avram, J. (2012). Emotional face processing in neurotypicals with autistic traits: implications for the broad autism phenotype. *Psychiatry Research, 198*(3), 489-494. https://doi.org/10.1016/j.psychres.2012.01.024

Molenberghs, P., Johnson, H., Henry, J. D., & Mattingley, J. B. (2016). Understanding the minds of others: A neuroimaging meta-analysis. *Neuroscience & Biobehavioral Reviews, 65,* 276-291. https://doi.org/10.1016/j.neubiorev.2016.03.020

Murphy, F. C., Nimmo-Smith, I., & Lawrence, A. D. (2003). Functional neuroanatomy of emotions: a meta-analysis. *Cognitive, Affective, & Behavioral Neuroscience, 3*(3), 207-233. https://doi.org/10.3758/CABN.3.3.207

O'Hare, A. E., Bremner, L., Nash, M., Happé, F., & Pettigrew, L. M. (2009). A clinical assessment tool for advanced theory of mind performance in 5 to 12 year olds. *Journal of Autism and Developmental Disorders, 39*(6), 916-928. https://doi.org/10.1007/s10803-009-0699-2

O'Nions, E., Sebastian, C. L., McCrory, E., Chantiluke, K., Happe, F., & Viding, E. (2014). Neural bases of Theory of Mind in children with autism spectrum disorders and children with conduct problems and callous-unemotional traits. *Developmental Science, 17*(5), 786-796. https://doi.org/10.1111/desc.12167

O'Reilly, K., Peterson, C. C., & Wellman, H. M. (2014). Sarcasm and advanced theory of mind understanding in children and adults with prelingual deafness. *Developmental Psychology, 50*(7), 1862. https://doi.org/10.1037/a0036654

Pardini, M., & Nichelli, P. F. (2009). Age-related decline in mentalizing skills across adult life span. *Experimental Aging Research, 35*(1), 98-106. https://doi.org/10.1080/03610730802545259

Parrila, R., Georgiou, G., & Corkett, J. (2007). University Students with a Significant Historyof Reading Difficulties: What Is and Is Not Compensated? *Exceptionality Education International, 17*(2), 195-220. https://doi.org/10.1037/t52486-000

Peñuelas-Calvo, I., Sareen, A., Sevilla-Llewellyn-Jones, J., & Fernández-Berrocal, P. (2019). The "Reading the Mind in the Eyes" Test in Autism-Spectrum Disorders Comparison with Healthy Controls: A Systematic Review and Meta-analysis. *Journal of Autism and Developmental Disorders, 49*(3), 1048-1061. https://doi.org/10.1007/s10803-018-3814-4

Perner, J. (1991). *Understanding the representational mind*: The MIT Press.

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

Petersen, R., Brakoulias, V., & Langdon, R. (2016). An experimental investigation of mentalization ability in borderline personality disorder. *Comprehensive Psychiatry, 64*, 12-21. https://doi.org/10.1016/j.comppsych.2015.10.004

Peterson, C., Slaughter, V., Moore, C., & Wellman, H. M. (2016). Peer social skills and theory of mind in children with autism, deafness, or typical development. *Developmental Psychology, 52*(1), 46-57. https://doi.org/10.1037/a0039833

Peterson, C. C., & Siegal, M. (1995). Deafness, conversation and theory of mind. *Journal of Child Psychology and Psychiatry, 36*(3), 459-474. https://doi.org/10.1111/j.1469-7610.1995.tb01303.x

Peterson, E., & Miller, S. (2012). The eyes test as a measure of individual differences: how much of the variance reflects verbal IQ? *Frontiers in Psychology, 3,* Article 220. https://doi.org/10.3389/fpsyg.2012.00220

Phillips, L. H., MacLean, R. D., & Allen, R. (2002). Age and the understanding of emotions: Neuropsychological and sociocognitive perspectives. *The Journals of Gerontology Series B: Psychological Sciences and Social Sciences, 57*(6), 526-530. https://doi.org/10.1093/geronb/57.6.P526

Pluta, A., Gawron, N., Sobańska, M., Wójcik, A. D., & Łojek, E. (2017). The nature of the relationship between neurocognition and theory of mind impairments in stroke patients. *Neuropsychology, 31*(6), 666-681. https://doi.org/10.1037/neu0000379

Preckel, K., Kanske, P., & Singer, T. (2018). On the interaction of social affect and cognition: empathy, compassion and theory of mind. *Current Opinion in Behavioral Sciences, 19,* 1-6. https://doi.org/10.1016/j.cobeha.2017.07.010

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

Preti, A., Vellante, M., & Petretto, D. R. (2017). The psychometric properties of the "Reading the Mind in the Eyes" Test: an item response theory (IRT) analysis. *Cognitive Neuropsychiatry, 22*(3), 233-253. https://doi.org/10.1080/13546805.2017.1300091

Purcell, A. L., Phillips, M., & Gruber, J. (2013). In your eyes: does theory of mind predict impaired life functioning in bipolar disorder? *Journal of Affective Disorders, 151*(3), 1113-1119. https://doi.org/10.1016/j.jad.2013.06.051

Rahman, M. S. (2017). The Advantages and Disadvantages of Using Qualitative and Quantitative Approaches and Methods in Language "Testing and Assessment" Research: A Literature Review. *Journal of Education and Learning, 6*(1), 102-112. https://doi.org/10.5539/jel.v6n1p102

Ratcliffe, M. (2006). Folk psychology' is not folk psychology. *Phenomenology and the Cognitive Sciences, 5*(1), 31-52. https://doi.org/10.1007/s11097-005-9010-y

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

Redondo, I., & Herrero-Fernández, D. (2018). Validation of the Reading the Mind in the Eyes Test in a healthy Spanish sample and women with anorexia nervosa. *Cognitive Neuropsychiatry, 23*(4), 201-217. https://doi.org/10.1080/13546805.2018.1461618

Reiter, A. M., Kanske, P., Eppinger, B., & Li, S. C. (2017). The aging of the social mind-differential effects on components of social understanding. *Scientific reports, 7*(1), 1-8. https://doi.org/10.1038/s41598-017-10669-4

Rice, K., & Redcay, E. (2014). 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

Ricker, T. J., & Sandry, J. (2018). The relationship between masking and short-term consolidation during recall from visual working memory. *Annals of the New York Academy of Sciences, 1424*(1), 91-101. https://doi.org/10.1111/nyas.13641

Rnic, K., Sabbagh, M. A., Washburn, D., Bagby, R. M., Ravindran, A., Kennedy, J. L., Strauss, J., & Harkness, K. L. (2018). Childhood emotional abuse, physical abuse, and neglect are associated with theory of mind decoding accuracy in young adults with depression. *Psychiatry Research, 268*, 501-507. https://doi.org/10.1016/j.psychres.2018.07.045

Sanvicente-Vieira, B., Kluwe-Schiavon, B., Wearick-Silva, L. E., Piccoli, G. L., Scherer, L., Tonelli, H. A., & Grassi-Oliveira, R. (2014). Revised reading the mind in the eyes test (RMET)-Brazilian version. *Brazilian Journal of Psychiatry, 36*(1), 60-67. https://doi.org/10.1590/1516-4446-2013-1162

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

Scheeren, A. M., de Rosnay, M., Koot, H. M., & Begeer, S. (2013). Rethinking theory of mind in high-functioning autism spectrum disorder. *Journal of Child Psychology and Psychiatry, 54*(6), 628-635. https://doi.org/10.1111/jcpp.12007

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

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

Shamay-Tsoory, S., Harari, H., Szepsenwol, O., & Levkovitz, Y. (2009). Neuropsychological evidence of impaired cognitive empathy in euthymic bipolar disorder. *The Journal of Neuropsychiatry and Clinical Neurosciences, 21*(1), 59-67. https://doi.org/10.1176/jnp.2009.21.1.59

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

Shamay-Tsoory, S. G., Tibi-Elhanany, Y., & Aharon-Peretz, J. (2006). The ventromedial prefrontal cortex is involved in understanding affective but not cognitive theory of mind stories. *Social Neuroscience, 1*(3-4), 149-166. https://doi.org/10.1080/17470910600985589

Slaughter, V., Imuta, K., Peterson, C. C., & Henry, J. D. (2015). Meta-analysis of theory of mind and peer popularity in the preschool and early school years. *Child Development, 86*(4), 1159-1174. https://doi.org/10.1111/cdev.12372

Söderstrand, P., & Almkvist, O. (2012). Psychometric data on the Eyes Test, the Faux Pas Test, and the Dewey Social Stories Test in a population-based Swedish adult sample. *Nordic Psychology, 64*(1), 30-43. https://doi.org/10.1080/19012276.2012.693729

Spanoudis, G. (2016). Theory of mind and specific language impairment in school-age children. *Journal of Communication Disorders, 61*, 83-96. https://doi.org/10.1016/j.jcomdis.2016.04.003

Spreng, R. N., McKinnon, M. C., Mar, R. A., & Levine, B. (2009). The Toronto Empathy Questionnaire: Scale development and initial validation of a factor-analytic solution to

multiple empathy measures. *Journal of Personality Assessment, 91*(1), 62-71. https://doi.org/10.1080/00223890802484381

Sprong, M., Schothorst, P., Vos, E., Hox, J., & Van Engeland, H. (2007). Theory of mind in schizophrenia: meta-analysis. *The British Journal of Psychiatry, 191*(1), 5-13. https://doi.org/10.1192/bjp.bp.107.035899

Stewart, E., Catroppa, C., & Lah, S. (2016). Theory of mind in patients with epilepsy: a systematic review and meta-analysis. *Neuropsychology Review, 26*(1), 3-24. https://doi.org/10.1007/s11065-015-9313-x

Stone, V. E., Baron-Cohen, S., & Knight, R. T. (1998). Frontal lobe contributions to theory of mind. *Journal of Cognitive Neuroscience, 10*(5), 640-656. https://doi.org/10.1162/089892998562942

Strauss, A., & Corbin, J. (1990). *Basics of qualitative research*. Sage publications.

Tager-Flusberg, H., & Sullivan, K. (2000). A componential view of theory of mind: evidence from Williams syndrome. *Cognition, 76*(1), 59-90. https://doi.org/10.1016/S0010-0277(00)00069-X

Tavakol, M., & Dennick, R. (2011). Making sense of Cronbach's alpha. *International Journal of Medical Education, 2*, 53-55. https://doi.org/10.5116/ijme.4dfb.8dfd

Taylor, M., Sachet, A. B., Maring, B. L., & Mannering, A. M. (2013). The assessment of elaborated role-play in young children: Invisible friends, personified objects, and pretend identities. *Social Development, 22*(1), 75-93. https://doi.org/10.1111/sode.12011

Thorne, S. (2000). Data analysis in qualitative research. *Evidence-based Nursing, 3*(3), 68-70. https://doi.org/10.1136/ebn.3.3.68

Thye, M. D., Murdaugh, D. L., & Kana, R. K. (2018). Brain mechanisms underlying reading the mind from eyes, voice, and actions. *Neuroscience, 374*, 172-186. https://doi.org/10.1016/j.neuroscience.2018.01.045

Trent, N. L., Park, C., Bercovitz, K., & Chapman, I. M. (2016). Trait socio-cognitive mindfulness is related to affective and cognitive empathy. *Journal of Adult Development, 23*(1), 62-67. https://doi.org/10.1007/s10804-015-9225-2

Turkstra, L. S. (2008). Theory of mind and use of cognitive state terms by adolescents with traumatic brain injury. *Aphasiology, 22*(10), 1054-1070. https://doi.org/10.1080/02687030701632187

Turner, R., & Felisberti, F. M. (2017). Measuring mindreading: a review of behavioral approaches to testing cognitive and affective mental state attribution in neurologically typical adults. *Frontiers in Psychology, 8*, 47. https://doi.org/10.3389/fpsyg.2017.00047

Ullman, M. T., & Pullman, M. Y. (2015). A compensatory role for declarative memory in neurodevelopmental disorders. *Neuroscience & Biobehavioral Reviews, 51*, 205-222. https://doi.org/10.1016/j.neubiorev.2015.01.008

Valle, A., Massaro, D., Castelli, I., & Marchetti, A. (2015). Theory of mind development in adolescence and early adulthood: the growing complexity of recursive thinking ability. *Europe's Journal of Psychology, 11*(1), 112. https://doi.org/10.5964/ejop.v11i1.829

Van der Meulen, A., Roerig, S., de Ruyter, D., van Lier, P., & Krabbendam, L. (2017). A comparison of children's ability to read children's and adults' mental states in an adaptation of the reading the mind in the eyes task. *Frontiers in Psychology, 8*, 594. https://doi.org/10.3389/fpsyg.2017.00594

Varnet, L., Meunier, F., Trollé, G., & Hoen, M. (2016). Direct viewing of dyslexics' compensatory strategies in speech in noise using auditory classification images. *PloS One, 11*(4), e0153781. https://doi.org/10.1371/journal.pone.0153781

Veissière, S. P., Constant, A., Ramstead, M. J., Friston, K. J., & Kirmayer, L. J. (2019). Thinking Through Other Minds: A Variational Approach to Cognition and Culture. *Behavioral and Brain Sciences,* 1-97. https://doi.org/10.1017/S0140525X19001213

Vésteinsdóttir, V., Reips, U. D., Joinson, A., & Thorsdottir, F. (2017). An item level evaluation of the Marlowe-Crowne Social Desirability Scale using item response theory on Icelandic Internet panel data and cognitive interviews. *Personality and Individual Differences, 107,* 164-173. https://doi.org/10.1016/j.paid.2016.11.023

Völlm, B. A., Taylor, A. N., Richardson, P., Corcoran, R., Stirling, J., McKie, S., Deakin, J. F., & Elliott, R. (2006). Neuronal correlates of theory of mind and empathy: a functional magnetic resonance imaging study in a nonverbal task. *Neuroimage, 29*(1), 90-98. https://doi.org/10.1016/j.neuroimage.2005.07.022

Vonk, J., & Pitzen, J. (2017). Believing in other minds: Accurate mentalizing does not predict religiosity. *Personality and Individual Differences, 115,* 70-76. https://doi.org/10.1016/j.paid.2016.06.008

Vonk, J., Zeigler-Hill, V., Ewing, D., Mercer, S., & Noser, A. E. (2015). Mindreading in the dark: Dark personality features and theory of mind. *Personality and Individual Differences, 87,* 50-54. https://doi.org/10.1016/j.paid.2015.07.025

Wade, M., Prime, H., Jenkins, J. M., Yeates, K. O., Williams, T., & Lee, K. (2018). On the relation between theory of mind and executive functioning: A developmental cognitive neuroscience perspective. *Psychonomic Bulletin & Review, 25*(6), 2119-2140. https://doi.org/10.3758/s13423-018-1459-0

Wakabayashi, A., & Katsumata, A. (2011). The Motion Picture Mind-Reading Test: Measuring Individual Differences of Social Cognitive Ability in a Young Adult Population in Japan. *Journal of Individual Differences, 32*(2), 55-64. https://doi.org/10.1016/j.rasd.2010.06.022

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

Wellman, H. M. (2014). *Making minds: How theory of mind develops*: Oxford University Press.

Wellman, H. M. (2018a). Theory of mind across the lifespan? *Zeitschrift für Psychologie, 226*(2), 136-138. https://doi.org/10.1027/2151-2604/a000330

Wellman, H. M. (2018b). 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., Cross, D., & Watson, J. (2001). Meta-analysis of theory-of-mind development: The truth about false belief. *Child Development, 72*(3), 655-684. https://doi.org/10.1111/1467-8624.00304

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

Wellman, H. M., & 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

Westerhof-Evers, H. J., Visser-Keizer, A. C., McDonald, S., & Spikman, J. M. (2014). Performance of healthy subjects on an ecologically valid test for social cognition: the short, Dutch Version of The Awareness of Social Inference Test (TASIT). *Journal of Clinical and Experimental Neuropsychology, 36*(10), 1031-1041. https://doi.org/10.1080/13803395.2014.966661

Westra, E. (2017). Spontaneous mindreading: A problem for the two-systems account. *Synthese, 194*(11), 4559-4581. https://doi.org/10.1007/s11229-016-1159-0

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

White, S. J., Coniston, D., Rogers, R., & Frith, U. (2011). Developing the Frith-Happé animations: A quick and objective test of Theory of Mind for adults with autism. *Autism Research, 4*(2), 149-154. https://doi.org/10.1002/aur.174

White, S. J., Frith, U., Rellecke, J., Al-Noor, Z., & Gilbert, S. J. (2014). Autistic adolescents show atypical activation of the brain′ s mentalizing system even without a prior history of mentalizing problems. *Neuropsychologia, 56*, 17-25. https://doi.org/10.1016/j.neuropsychologia.2013.12.013

Woody, R. (2015). *Posttraumatic Stress and Emotion Recognition* [Doctoral dissertation, University of Arkansas] UARKive http://hdl.handle.net/10826/1174

Zhang, D., Pang, Y., Cai, W., Fazio, R. L., Ge, J., Su, Q., Xu, S., Pan, Y., Chen, S., & Zhang, H. (2016). Development and psychometric properties of an informant assessment

scale of theory of mind for adults with traumatic brain injury. *Neuropsychological Rehabilitation, 26*(4), 481-501. https://doi.org/10.1080/09602011.2015.1030431

Appendices

Appendix A

Reading the Mind in the Eyes - Revised test materials

Instructions:

In the upcoming screens you will be presented with a series of pictures of the eye regions of different individuals along with a variety of mental state terms.

For each set of eyes, please select using number keys 1, 2, 3 or 4 the word that best describes what the person in the picture is thinking or feeling.

You may feel that more than one word is applicable but please choose just the word which you consider to be most suitable. And make sure to read every word before making your choice.

Three variations of this section exist, depending on which variation the participant has received:

20 second variation: In this task you will have 20 seconds to answer each question. OR

5 second variation: In this task you will have 5 seconds to answer each question. OR

Occlusion variation: In this task you will see each picture briefly (0.5 seconds), after which you will have 5 seconds to answer each question.

When you answer, **use the number keys (1-4) at the top of your keyboard** - not those on the separate numeric keypad (if you have one).

If you run out of time the trial will end and you will be taken to the next picture.

RMET-R Pictures:

Practice question

Question 1

Question 2

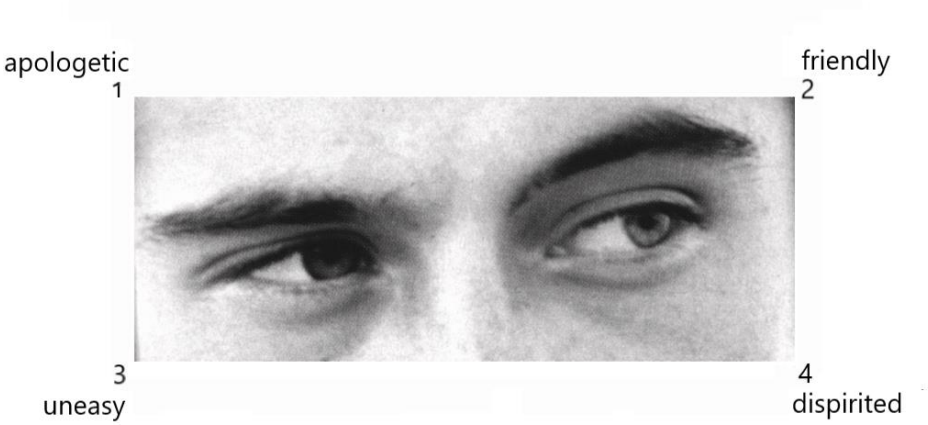
Question 3

Question 4

Question 5

Question 6

Question 7

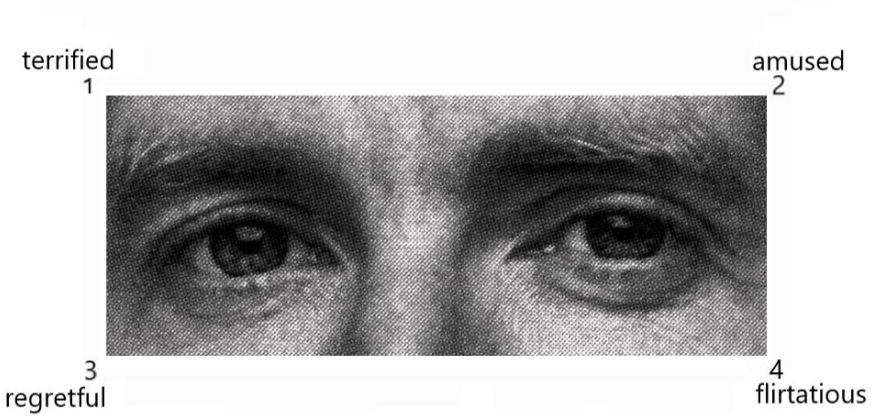


Question 8

Question 9

Question 10

Question 11



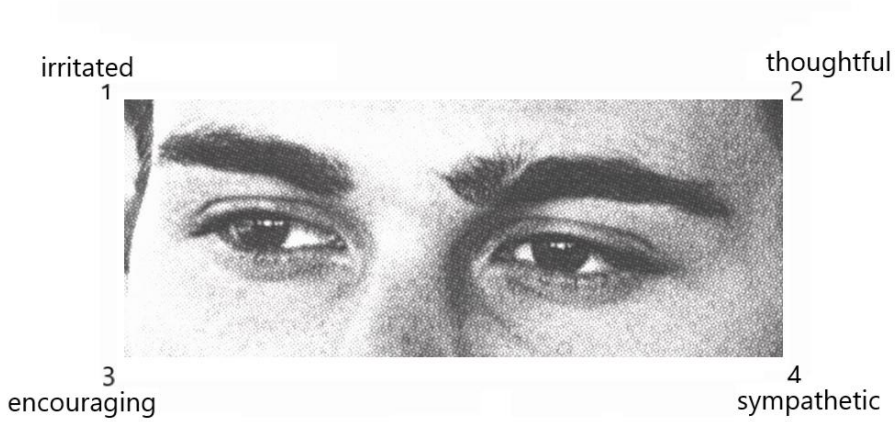
Question 12

Question 13

Question 14

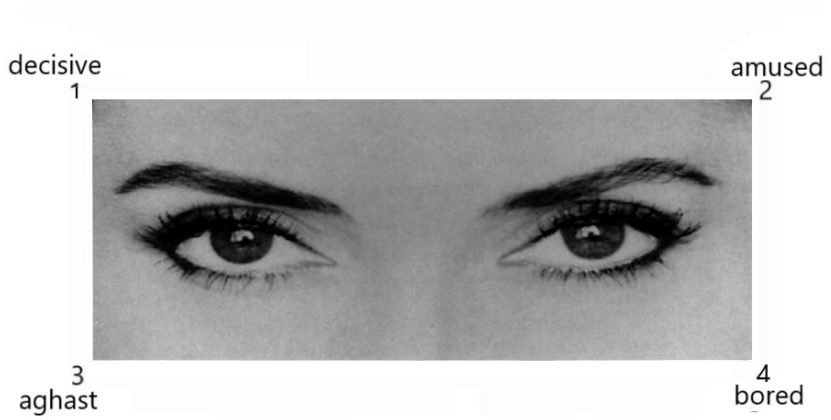
Question 15

Question 16

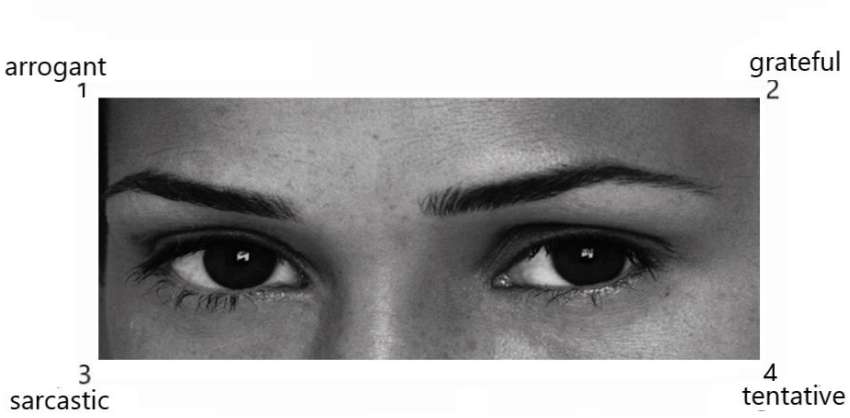


Question 17

Question 18



Question 19

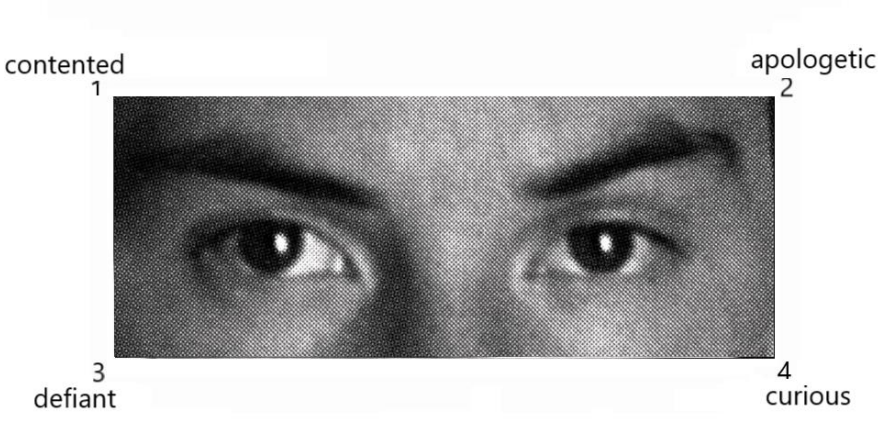


Question 20

Question 21

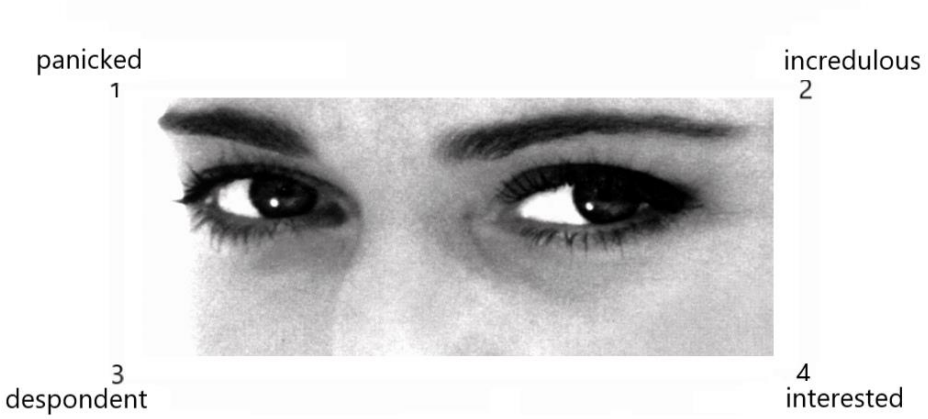
Question 22

Question 23

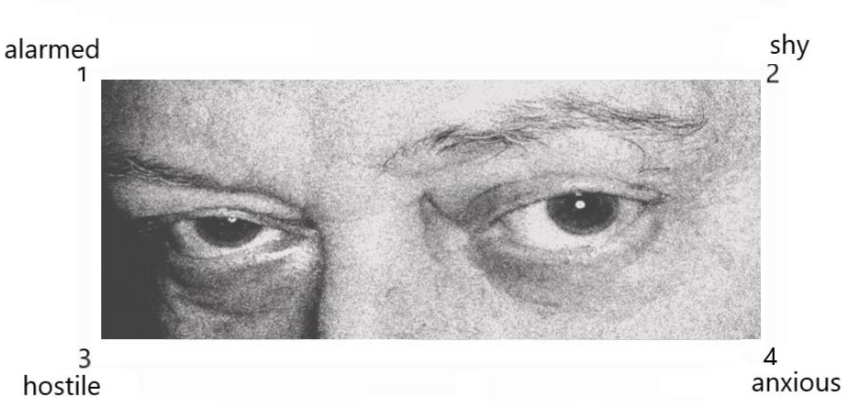


Question 24

Question 25



Question 26



Question 27

Question 28

Question 29

Question 30

Question 31

Question 32

Question 33

Question 34

Question 35

Question 36

RMET-R answers

Note: **correct answer**

Practice: jealous	**panicked**	arrogant	hateful
1:	**playful**	comforting	irritated	bored
2:	terrified	**upset**	arrogant	annoyed
3:	joking	flustered	**desire**	convinced
4:	joking	**insisting**	amused	relaxed
5:	irritated	sarcastic	**worried**	friendly
6:	aghast	**fantasizing**	impatient	alarmed
7:	apologetic	friendly	**uneasy**	dispirited
8:	**despondent**	relieved	shy	excited
9:	annoyed	hostile	horrified	**preoccupied**
10:	**cautious**	insisting	bored	aghast
11:	terrified	amused	**regretful**	flirtatious
12:	indifferent	embarrassed	**sceptical**	dispirited
13:	decisive	**anticipating**	threatening	shy
14:	irritated	disappointed	depressed	**accusing**
15:	**contemplative**	flustered	encouraging	amused
16:	irritated	**thoughtful**	encouraging	sympathetic
17:	**doubtful**	affectionate	playful	aghast
18:	**decisive**	amused	aghast	bored
19:	arrogant	grateful	sarcastic	**tentative**
20:	dominant	**friendly**	guilty	horrified
21:	embarrassed	**fantasizing**	confused	panicked
22:	**preoccupied**	grateful	insisting	imploring

23:	contented	apologetic	**defiant**	curious
24:	**pensive**	irritated	excited	hostile
25:	panicked	incredulous	despondent	**interested**
26:	alarmed	shy	**hostile**	anxious
27:	joking	**cautious**	arrogant	reassuring
28:	**interested**	joking	affectionate	contented
29:	impatient	aghast	irritated	**reflective**
30:	grateful	**flirtatious**	hostile	disappointed
31:	ashamed	**confident**	joking	dispirited
32:	**serious**	ashamed	bewildered	alarmed
33:	embarrassed	guilty	fantasizing	**concerned**
34:	aghast	baffled	**distrustful**	terrified
35:	puzzled	**nervous**	insisting	contemplative
36:	ashamed	nervous	**suspicious**	Indecisive

Appendix B

RMET-R mental state terms with accompanying definition and example

RMET-R mental state glossary taken from Baron-Cohen et al., 2001

ACCUSING	blaming The policeman was accusing the man of stealing a wallet.
AFFECTIONATE	showing fondness towards someone Most mothers are affectionate to their babies by giving them lots of kisses and cuddles.
AGHAST	horrified, astonished, alarmed Jane was aghast when she discovered her house had been burgled.
ALARMED	fearful, worried, filled with anxiety Claire was alarmed when she thought she was being followed home.
AMUSED	finding something funny I was amused by a funny joke someone told me.
ANNOYED	irritated, displeased Jack was annoyed when he found out he had missed the last bus home.
ANTICIPATING	expecting At the start of the football match, the fans were anticipating a quick goal.
ANXIOUS	worried, tense, uneasy The student was feeling anxious before taking her final exams.
APOLOGETIC	feeling sorry The waiter was very apologetic when he spilt soup all over the customer.
ARROGANT	conceited, self-important, having a big opinion of oneself The arrogant man thought he knew more about politics than everyone else in the room.
ASHAMED	overcome with shame or guilt The boy felt ashamed when his mother discovered him stealing money from her purse.
ASSERTIVE	confident, dominant, sure of oneself The assertive woman demanded that the shop give her a refund.
BAFFLED	confused, puzzled, dumbfounded The detectives were completely baffled by the murder case.
BEWILDERED	utterly confused, puzzled, dazed The child was bewildered when visiting the big city for the first time.
CAUTIOUS	careful, wary

Sarah was always a bit cautious when talking to someone she did not know.

COMFORTING consoling, compassionate
The nurse was comforting the wounded soldier.

CONCERNED worried, troubled
The doctor was concerned when his patient took a turn for the worse.

CONFIDENT self-assured, believing in oneself
The tennis player was feeling very confident about winning his match.

CONFUSED puzzled, perplexed
Lizzie was so confused by the directions given to her, she got lost.

CONTEMPLATIVE reflective, thoughtful, considering
John was in a contemplative mood on the eve of his 60th birthday.

CONTENTED satisfied
After a nice walk and a good meal, David felt very contented.

CONVINCED certain, absolutely positive
Richard was convinced he had come to the right decision.

CURIOUS inquisitive, inquiring, prying
Louise was curious about the strange shaped parcel.

DECIDING making your mind up
The man was deciding whom to vote for in the election.

DECISIVE already made your mind up
Jane looked very decisive as she walked into the polling station.

DEFIANT insolent, bold, don’t care what anyone else thinks
The animal protester remained defiant even after being sent to prison.

DEPRESSED miserable
George was depressed when he didn't receive any birthday cards.

DESIRE passion, lust, longing for
Kate had a strong desire for chocolate.

DESPONDENT gloomy, despairing, without hope
Gary was despondent when he did not get the job he wanted.

DISAPPOINTED displeased, disgruntled
Manchester United fans were disappointed not to win the Championship.

DISPIRITED glum, miserable, low
Adam was dispirited when he failed his exams.

DISTRUSTFUL suspicious, doubtful, wary

The old woman was distrustful of the stranger at her door.

DOMINANT commanding, bossy
The sergeant major looked dominant as he inspected the new recruits.

DOUBTFUL dubious, suspicious, not really believing
Mary was doubtful that her son was telling the truth.

DUBIOUS doubtful, suspicious
Peter was dubious when offered a surprisingly cheap television in a pub.

EAGER keen
On Christmas morning, the children were eager to open their presents.

EARNEST having a serious intention
Harry was very earnest about his religious beliefs.

EMBARRASSED ashamed
After forgetting a colleague's name, Jenny felt very embarrassed.

ENCOURAGING hopeful, heartening, supporting
All the parents were encouraging their children in the school sports day.

ENTERTAINED absorbed and amused or pleased by something
I was very entertained by the magician.

ENTHUSIASTIC very eager, keen
Susan felt very enthusiastic about her new fitness plan.

FANTASIZING daydreaming
Emma was fantasizing about being a film star.

FASCINATED captivated, really interested
At the seaside, the children were fascinated by the creatures in the rock pools.

FEARFUL terrified, worried
In the dark streets, the women felt fearful.

FLIRTATIOUS brazen, saucy, teasing, playful
Connie was accused of being flirtatious when she winked at a stranger at a party.

FLUSTERED confused, nervous and upset
Sarah felt a bit flustered when she realised how late she was for the meeting and that she had forgotten an important document.

FRIENDLY sociable, amiable
The friendly girl showed the tourists the way to the town centre.

GRATEFUL thankful
Kelly was very grateful for the kindness shown by the stranger.

GUILTY feeling sorry for doing something wrong
Charlie felt guilty about having an affair.

HATEFUL showing intense dislike
The two sisters were hateful to each other and always fighting.

HOPEFUL optimistic
Larry was hopeful that the post would bring good news.

HORRIFIED terrified, appalled
The man was horrified to discover that his new wife was already married.

HOSTILE unfriendly
The two neighbours were hostile towards each other because of an argument about loud music.

IMPATIENT restless, wanting something to happen soon
Jane grew increasingly impatient as she waited for her friend who was already 20 minutes late.

IMPLORING begging, pleading
Nicola looked imploring as she tried to persuade her dad to lend her the car.

INCREDULOUS not believing
Simon was incredulous when he heard that he had won the lottery.

INDECISIVE unsure, hesitant, unable to make your mind up
Tammy was so indecisive that she couldn't even decide what to have for lunch.

INDIFFERENT disinterested, unresponsive, don't care
Terry was completely indifferent as to whether they went to the cinema or the pub.

INSISTING demanding, persisting, maintaining
After a work outing, Frank was insisting he paid the bill for everyone.

INSULTING rude, offensive
The football crowd was insulting the referee after he gave a penalty.

INTERESTED inquiring, curious
After seeing Jurassic Park, Hugh grew very interested in dinosaurs.

INTRIGUED very curious, very interested
A mystery phone call intrigued Zoe.

IRRITATED exasperated, annoyed
Frances was irritated by all the junk mail she received.

JEALOUS envious
Tony was jealous of all the taller, better-looking boys in his class.

JOKING being funny, playful

Gary was always joking with his friends.

NERVOUS apprehensive, tense, worried
Just before her job interview, Alice felt very nervous.

OFFENDED insulted, wounded, having hurt feelings
When someone made a joke about her weight, Martha felt very offended.

PANICKED distraught, feeling of terror or anxiety
On waking to find the house on fire, the whole family was panicked.

PENSIVE thinking about something slightly worrying
Susie looked pensive on the way to meeting her boyfriend's parents for the first time.

PERPLEXED bewildered, puzzled, confused
Frank was perplexed by the disappearance of his garden gnomes.

PLAYFUL full of high spirits and fun
Neil was feeling playful at his birthday party.

PREOCCUPIED absorbed, engrossed in one's own thoughts
Worrying about her mother's illness made Debbie preoccupied at work

PUZZLED perplexed, bewildered, confused
After doing the crossword for an hour, June was still puzzled by one clue.

REASSURING supporting, encouraging, giving someone confidence
Andy tried to look reassuring as he told his wife that her new dress did suit her.

REFLECTIVE contemplative, thoughtful
George was in a reflective mood as he thought about what he'd done with his life.

REGRETFUL sorry
Lee was always regretful that he had never travelled when he was younger.

RELAXED taking it easy, calm, carefree
On holiday, Pam felt happy and relaxed.

RELIEVED freed from worry or anxiety
At the restaurant, Ray was relieved to find that he had not forgotten his wallet.

RESENTFUL bitter, hostile
The businessman felt very resentful towards his younger colleague who had been promoted above him.

SARCASTIC cynical, mocking, scornful
The comedian made a sarcastic comment when someone came into the theatre late.

SATISFIED content, fulfilled
Steve felt very satisfied after he had got his new flat just how he wanted it.

SCEPTICAL	doubtful, suspicious, mistrusting Patrick looked sceptical as someone read out his horoscope to him.
SERIOUS	solemn, grave The bank manager looked serious as he refused Nigel an overdraft.
STERN	severe, strict, firm The teacher looked very stern as he told the class off.
SUSPICIOUS	disbelieving, suspecting, doubting After Sam had lost his wallet for the second time at work, he grew suspicious of one of his colleagues.
SYMPATHETIC	kind, compassionate The nurse looked sympathetic as she told the patient the bad news.
TENTATIVE	hesitant, uncertain, cautious Andrew felt a bit tentative as he went into the room full of strangers.
TERRIFIED	alarmed, fearful The boy was terrified when he thought he saw a ghost.
THOUGHTFUL	thinking about something Phil looked thoughtful as he sat waiting for the girlfriend he was about to finish with.
THREATENING	menacing, intimidating The large, drunken man was acting in a very threatening way.
UNEASY	unsettled, apprehensive, troubled Karen felt slightly uneasy about accepting a lift from the man she had only met that day.
UPSET	agitated, worried, uneasy The man was very upset when his mother died.
WORRIED	anxious, fretful, troubled When her cat went missing, the girl was very worried

Appendix C

Manipulation check questions

The following questions ask about the initial eye impressions task that you just completed.

Q1. How difficult did you find the eye impressions task?

Paired with a 10-point answer scale ranging from: 1: Extremely easy to 10: Extremely hard

Q2. How much mental effort was required for the task?

Paired with a 10-point answer scale ranging from: 1: Extremely little effort to 10: Extremely large effort

Q3. How well do you think you did on the task?

Paired with a 10-point answer scale ranging from: 1: Extremely well to 10: Extremely poorly

Q4. I read instructions carefully. To show that you are reading these instructions, please leave this question blank.

Paired with a 10-point answer scale ranging from: 1: Extremely true to 10:Extremely false

Note: Q4 is one of the two attention checks used in this study and is not part of the Manipulation checks. Any answer on this question is considered an incorrect answer.

Appendix D

Toronto Empathy Questionnaire materials

Below is a list of statements. Please read each statement carefully and rate how frequently you feel or act in the manner described. There are no right or wrong answers or trick questions. Please answer each question as honestly as you can.

Q1. When someone else is feeling excited, I tend to get excited too.

Q2. Other people's misfortunes do not disturb me a great deal.

Q3. It upsets me to see someone being treated disrespectfully.

Q4. I remain unaffected when someone close to me is happy.

Q5. I enjoy making other people feel better.

Q6. I have tender, concerned feelings for people less fortunate than me.

Q7. When a friend starts to talk about his\her problems, I try to steer the conversation towards something else.

Q8. I can tell when others are sad even when they do not say anything.

Q9. I find that I am "in tune" with other people's moods.

Q10. I do not feel sympathy for people who cause their own serious illnesses.

Q11. I become irritated when someone cries.

Q12. I am not really interested in how other people feel.

Q13. I get a strong urge to help when I see someone who is upset

Q14. When I see someone being treated unfairly, I do not feel very much pity for them.

Q15. I find it silly for people to cry out of happiness.

Q16. When I see someone being taken advantage of, I feel kind of protective towards him\her.

5-point answer Scale: Never, Rarely, Sometimes, Often, Always

Scoring key:

Question: 1, 3, 5, 6, 8, 9, 13, 16. Never = 0; Rarely = 1; Sometimes = 2; Often = 3; Always = 4.
Question: 2, 4, 7, 10, 11, 12, 14, 15. Never = 4; Rarely = 3; Sometimes = 2; Often = 1; Always = 0.

Appendix E

Mentalization Scale (MentS) materials

Please read each of the following statements carefully. Using the selections available, indicate how correctly each statement describes you.

Q1. I find it important to understand reasons for my behaviour.

Q2. When I make conclusions about other people's personality traits I carefully observe what they say and do.

Q3. I can recognize other people's feelings.

Q4. I often think about other people and their behaviour.

Q5. Usually I can recognize what makes people feel uneasy.

Q6. I can sympathize with other people's feelings.

Q7. When someone annoys me I try to understand why I react in that way.

Q8. When I get upset I am not sure whether I am sad, afraid, or angry.

Q9. I do not like to waste time trying to understand in detail other people's behaviour.

Q10. I can make good predictions of other people's behaviour when I know their beliefs and feelings.

Q11. Often I cannot explain, even to myself, why I did something.

Q12. Sometimes I can understand someone's feelings before s/he tells me anything.

Q13. I find it important to understand what happens in my relationships with people close to me.

Q14. I do not want to find out something about myself that I will not like.

Q15. To understand someone's behaviour, we need to know her/his thoughts, wishes, and feelings.

Q16. I often talk about emotions with people that I am close to.

Q17. I like reading books and newspaper articles about psychological subjects.

Q18. I find it difficult to admit to myself that I am sad, hurt, or afraid.

Q19. I do not like to think about my problems.

Q20. I can describe significant traits of people who are close to me with precision and in detail.

Q21. I am often confused about my exact feelings.

Q22. It is difficult for me to find adequate words to express my feelings.

Q23. People tell me that I understand them and give them sound advice.

Q24. I have always been interested in why people behave in certain ways.

Q25. I can easily describe what I feel.

Q26. I once owned a three headed dog.

Q27. While people talk about their feelings and needs my thoughts often drift away.

Q28. Since we all depend on life circumstances, it is meaningless to think of other people's intentions or wishes.

Q29. One of the most important things that children should learn is to express their feelings and wishes.

5-point answer scale: Completely incorrect, Mostly incorrect, Both correct and incorrect, Mostly correct, Completely correct

Scoring key:

Completely incorrect: 1

Mostly incorrect: 2

Both correct and incorrect: 3

Mostly correct: 4

Completely correct: 5

MentS subscales

MentS-Self: 8, 11, 14, 18, 19, 21, 22, 27.

MentS-Other: 2, 3, 5, 6, 10, 12, 20, 23, 25, 29.

MentS-Motivation: 1, 4, 5, 7, 9, 13, 16, 17, 24, 28.

Note: Q26 is the 1st attention check question of this study and is not part of the MentS. The correct answer is completely incorrect

Appendix F

Marlowe-Crowne Social Desirability Scale Short-Form (MCSDS-SF) materials

Listed below are a number of statements concerning personal attitudes and traits. Read each item and decide whether the statement is true or false as it pertains to you personally.

Q1. I have never intensely disliked anyone.

Q2. I sometimes feel resentful when I don't get my way.

Q3. No matter who I'm talking to, I'm always a good listener.

Q4. There have been occasions when I took advantage of someone.

Q5. I'm always willing to admit it when I make a mistake.

Q6. I sometimes try to get even, rather than forgive and forget.

Q7. There have been occasions when I felt like smashing things.

Q8. There have been times when I was quite jealous of the good fortune of others.

Q9. I have never felt that I was punished without cause.

Q10. I have never deliberately said something that hurt someone's feelings.

Answer scale:

True / False

Scoring key:

1 point true for Questions: 1, 3, 5, 9, 10.

1 point false for questions: 2, 4, 6, 7, 8.

Appendix G

Self-report questionnaire of methods used to assess RMET-R stimuli

When people make quick impressions based only on seeing the eyes of another person they often use specific techniques for forming that impression quickly. In this study you have made many judgements about the mental states of the people depicted in the photographs.

What we would like now is for you to tell us what main method or technique you used to figure out which mental state best fit each picture of someone's eyes. People vary on how they do this, so note down how you went about making a decision in the space presented below.

A number of individuals also use multiple techniques to figure out the mental state, with some individuals switching strategies depending on difficulty or commonly using multiple strategies together to come to a decision.

Listed below are some of the most common strategies found to be used in previous experiments. Please indicate if you used any of the strategies below in addition to the main method you noted on the previous page.

- ☐ Intuition/Just instantly know
- ☐ Imagining the rest of the face (visual imagery)
- ☐ Comparing the eyes to previous experiences/events
- ☐ Rules of thumb (eye direction, level of squint, etc)
- ☐ Process of elimination through knowledge about the Mental state terms (rule out obviously wrong answers and make best guess)

In addition could you also indicate whether you typically switched between individual strategies depending on difficulty, commonly use a combination of strategies to reach an answer, or switched to a combination of strategies for more difficult judgements:

- ○ Only use single main strategy
- ○ Switch between individual strategies
- ○ Commonly use combination of strategies
- ○ Use combination of strategies for difficult judgements

Appendix H

Demographic questionnaire

Demographics

Please complete the questionnaire below then click the button below to continue.

Age (years): Space for answer

Gender:

○ Male ○ Female ○ Gender Diverse

Note: If Gender diverse is chosen an optional details textbox will also appear.

Are you a proficient English speaker?

○ Yes ○ No

Have you ever performed the reading the mind in the eyes test before?

○ Yes ○ No

Do you have any major eye sight problems?

○ Yes ○ No

Do you have any physical or mental difficulties that that may impede your reading abilities or comprehension?

○ Yes ○ No

What is your highest level of education you have obtained?

○ Early childhood education

○ Primary Education

○ Lower Secondary Education

○ Upper Secondary Education

○ Post-secondary non-Tertiary Education

○ Short-cycle tertiary education

○ Bachelors degree or equivalent tertiary education level

○ Masters degree or equivalent tertiary education level

○ Doctoral degree or equivalent tertiary education level

Have you ever been diagnosed as having of the following conditions?

☐ Autistic Spectrum Disorder (Or other related condition)

☐ Schizophrenia

☐ Traumatic Brain Injury

Appendix I

Main experiment information page

Welcome

You are invited to take part in an online psychology experiment. Before you decide whether to take part in this study please read the following information carefully.

Why are we doing this research?

This experiment contributes to a master's project being conducted in the School of Psychology at Massey University (NZ). The study investigates the effect of adding time limitations to a commonly used theory of mind task.

What does this experiment involve?

In this experiment you will complete some questions that ask you to reflect on your own mental states, dispositions and abilities. You will also perform a series of tasks in which you will see the eye region of a face and then select which out of a selection of terms best matches the mental state of the person in the picture. The experiment will end with some demographic questions and some reflections about how you completed the eye task. The experiment will take approximately 20 minutes to complete.

What will you do with the information I provide?

No identifiable information will be collected as part of this study, and all data will be amalgamated before it is analysed for research purposes and posted on a public research repository.

Are there any potential risks?

There are no foreseeable risks involved in this study and your participation is completely voluntary. Participants are free to stop participating at any time or decline to answer any question without penalty.

This project has been evaluated by peer review and judged to be low risk. Consequently, it has not been reviewed by one of the University's Human Ethics Committees. The researcher(s) named in this document are responsible for the ethical conduct of this research.

If you have any concerns about the conduct of this research that you want to raise with someone other than the researcher(s), please contact Professor Craig Johnson, Director (Research Ethics), email humanethics@massey.ac.nz.

Researcher contacts

If you have any questions or concerns about this research, please contact the primary researcher of this project; Liam Allen at liamresearch4@gmail.com. You may also contact the project's research supervisors: Dr Michael Philipp at m.philipp@massey.ac.nz or Dr Stephen Hill at S.R.Hill@massey.ac.nz.

Appendix J

Main experiment consent page

If you are interested to participate in this research, please check the boxes below confirming that you understand and agree with each statement.

- ☐ I confirm that I have read and understood the information given about the study on the previous page.
- ☐ I understand that my participation is voluntary and that I am free to stop participating at any time.
- ☐ I agree to have my anonymised data included in a public, online research data repository that will be available to other researchers.
- ☐ I am at least 18 years old.
- ☐ I agree to take part in this experiment.

Appendix K

Post-experiment debrief

Did you encounter any problems or notice any errors as you took part in this study?

If so, please let us know, as this can help us better understand the information you've shared with us.

Space to write down answers

Based on your experience during this study, what would you guess is the hypothesis of this research?

Please limit your response to 1 or 2 sentences.

Space to write down answers

Thank you for completing this study.

What was the aim of this study?

Over the two decades there has been increasing interest in understanding peoples' theory of mind. Theory of mind reflects our ability to make good guesses about what other people are thinking. Our study is particularly interested in understanding how we can improve our ability to measure theory of mind.

Our study used an adapted version of the Reading the Mind in The Eyes Test (RMET), developed by Baron-Cohen, Wheelwright, Hill, Raste and Plumb (2001) to measure theory of mind ability. In the version of the RMET that you took, we added a time limit for each trial to see if this affects the test's validity.

This experiment also used a measure of mentalising developed by Dimitrijević, Hanak, Altaras Dimitrijević, and Jolić Marjanović (2018) and a measure of empathy developed by Spreng, McKinnon, Mar and Levine (2009). These measures are included because they are often closely related to peoples' theory of mind abilities.

If you would like to know about the results of this study and our pilot studies, a summary of the results will start to be made available online at the link below within the next couple of weeks:

http://bit.ly/EyeImpressionResults

Further information

If you would like to learn more about theory of mind (ToM) a good place to start is: https://en.wikipedia.org/wiki/Theory_of_mind

If you would like to know more about the assessments used in this experiment, please see the below articles.

Reading the Mind in the Eyes test:
https://depts.washington.edu/uwcssc/sites/default/files/hw00/d40/uwcssc/sites/default/files/Mind%20in%20the%20Eyes%20Scale_0.pdf

Toronto Empathy Questionnaire:
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2775495/

Mentalization Scale:
https://www.tandfonline.com/doi/abs/10.1080/00223891.2017.1310730

Click here to return to prolific

Appendix L

Q-Q plots of standardised residuals of all variables used in One-way ANOVAs in this study

Figure L1

Q-Q plot of the standardized residuals of Difficulty

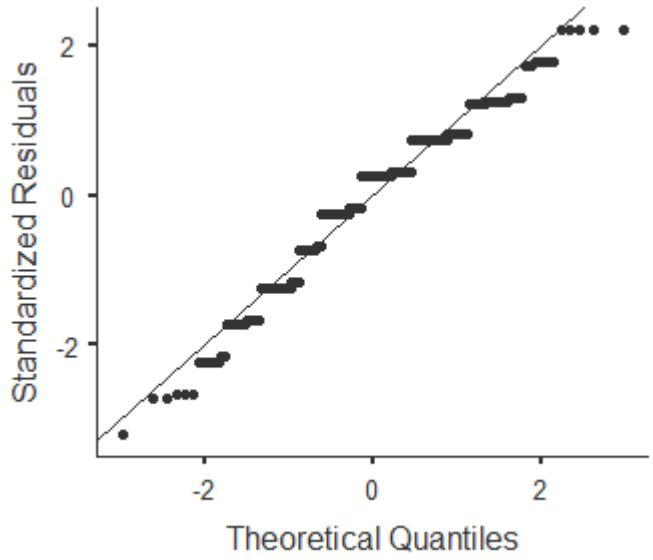


Figure L2

Q-Q plot of the standardized residuals of Mental effort

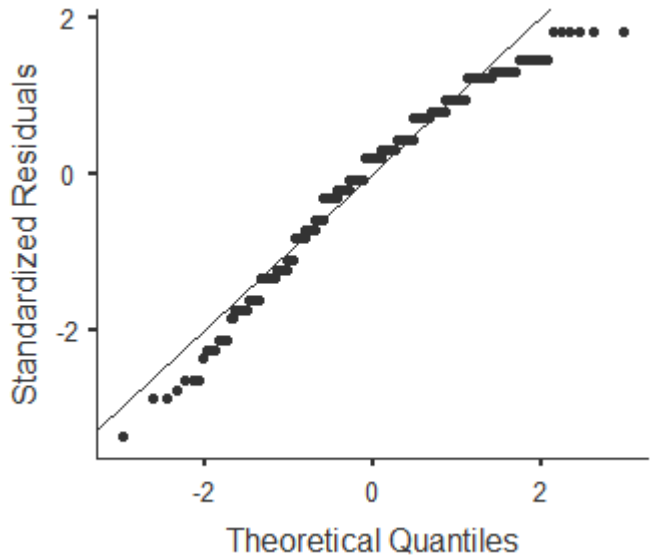


Figure L3

Q-Q plot of the standardized residuals of Believed performance

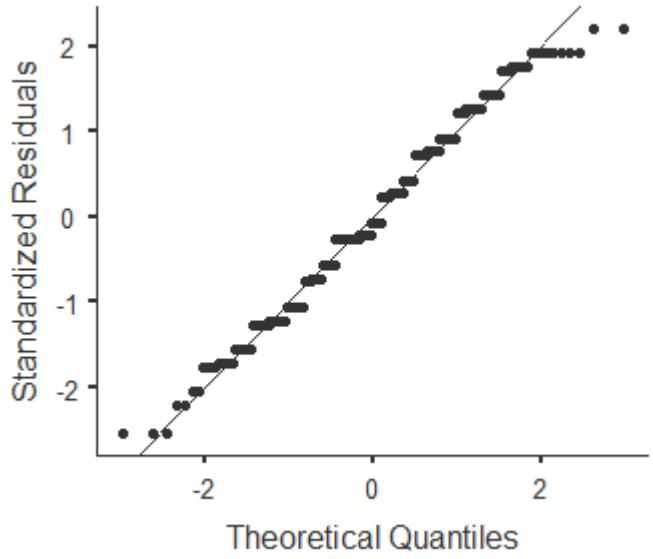


Figure L4

Q-Q plot of the standardized residuals of RMET-R scores

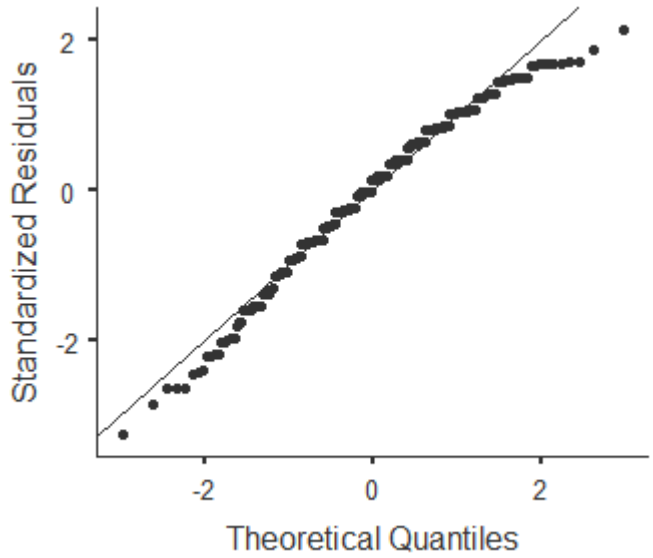


Figure L5

Q-Q plot of the standardized residuals of RMET-R Occluded variation scores

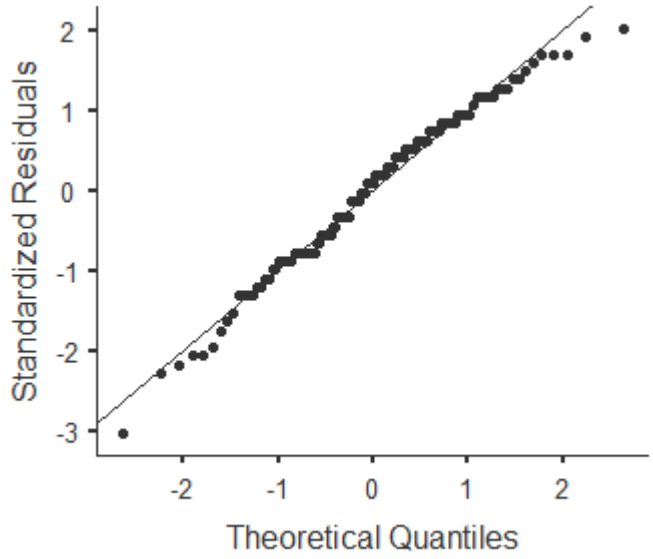


Figure L6

Q-Q plot of the standardized residuals of RMET-R Short variation scores

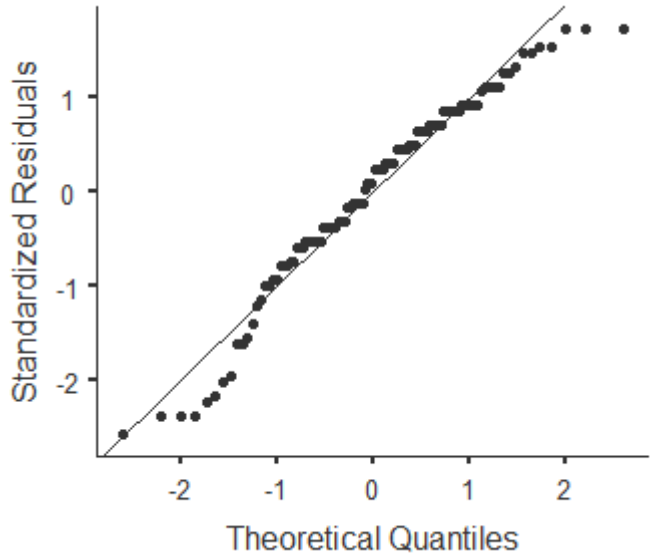


Figure L7

Q-Q plot of the standardized residuals of RMET-R Long variation scores

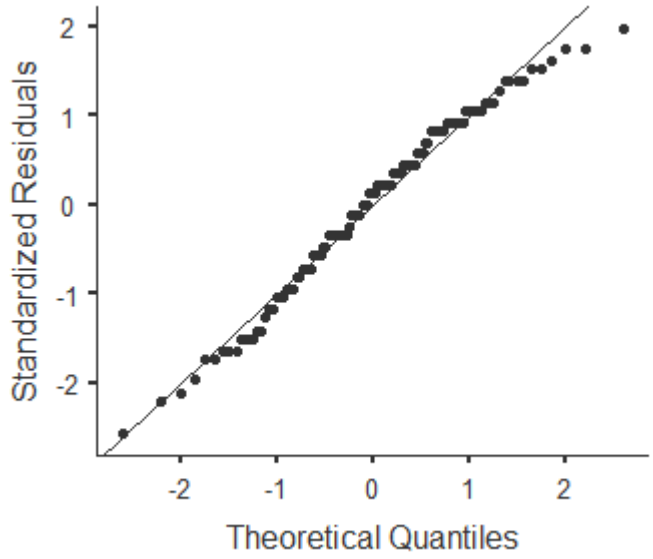


Figure L8

Q-Q plot of the standardized residuals of TEQ scores

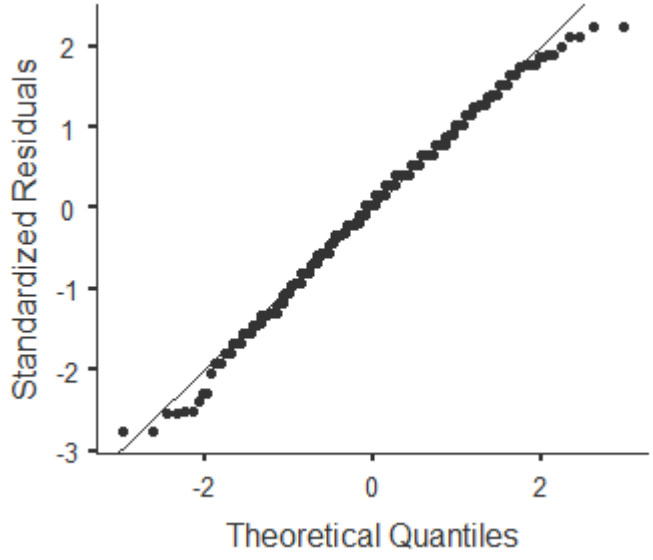


Figure L9

Q-Q plot of the standardized residuals of MentS scores

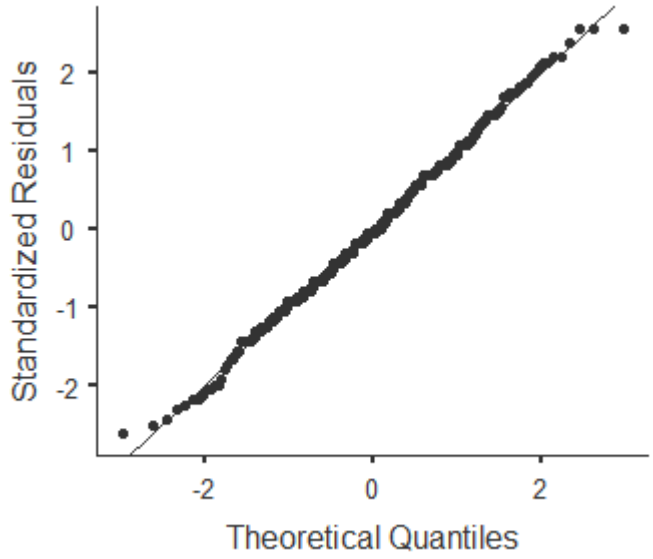


Figure L10

Q-Q plot of the standardized residuals of MentS Motivation scores

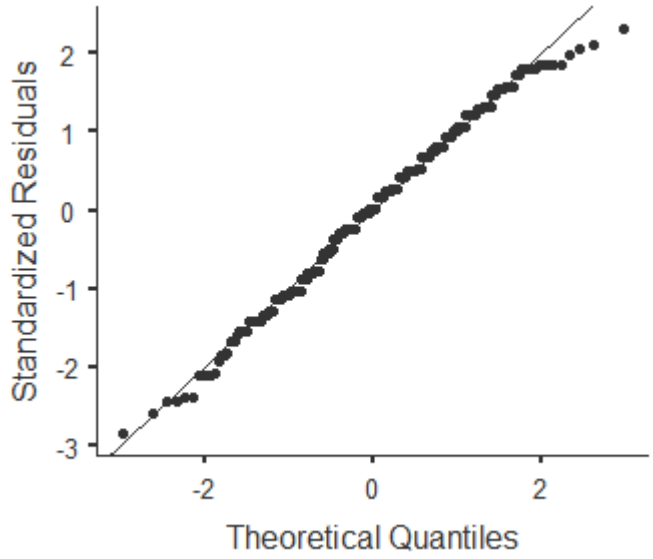


Figure L11

Q-Q plot of the standardized residuals of MentS Other scores

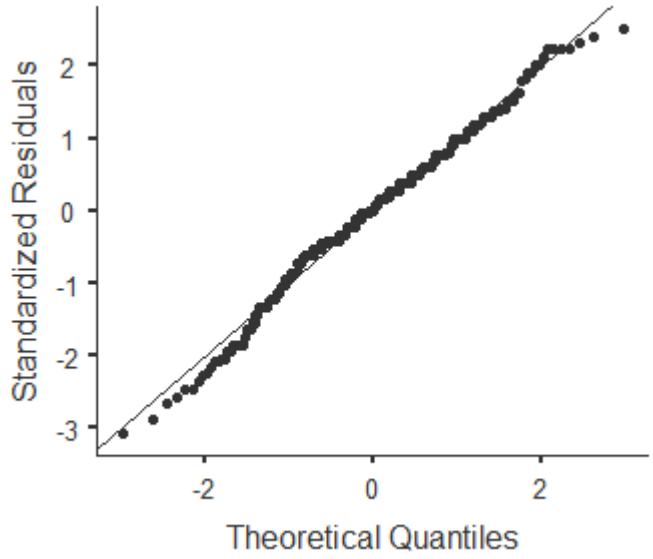


Figure L12

Q-Q plot of the standardized residuals of MentS Self scores

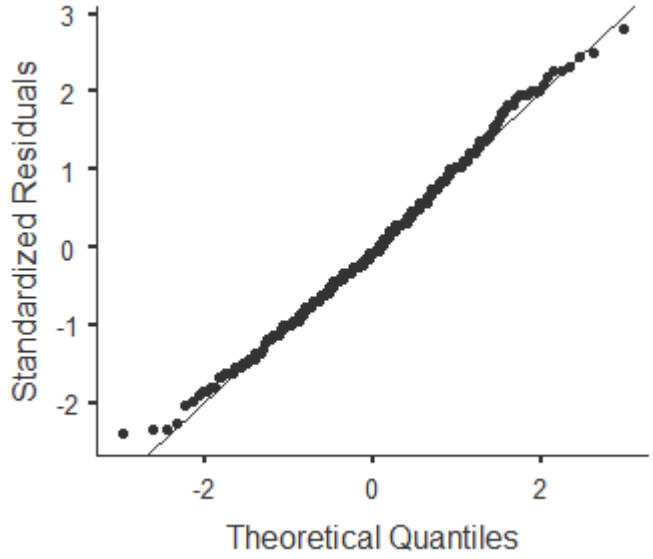
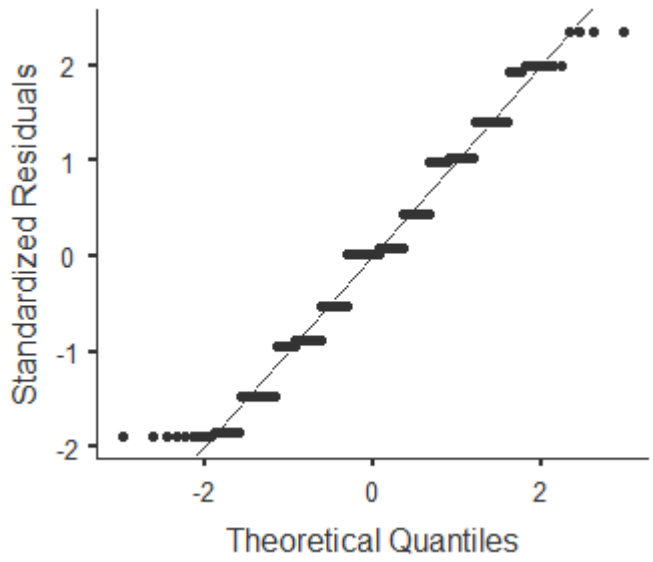


Figure L13

Q-Q plot of the standardized residuals of Total number of strategies



Appendix M

Scatterplots of the predicted residuals in each step of the regression analyses in this study

Figure M1

Scatterplot of the predicted residuals of step 1 of the TEQ-RMET-R Regression

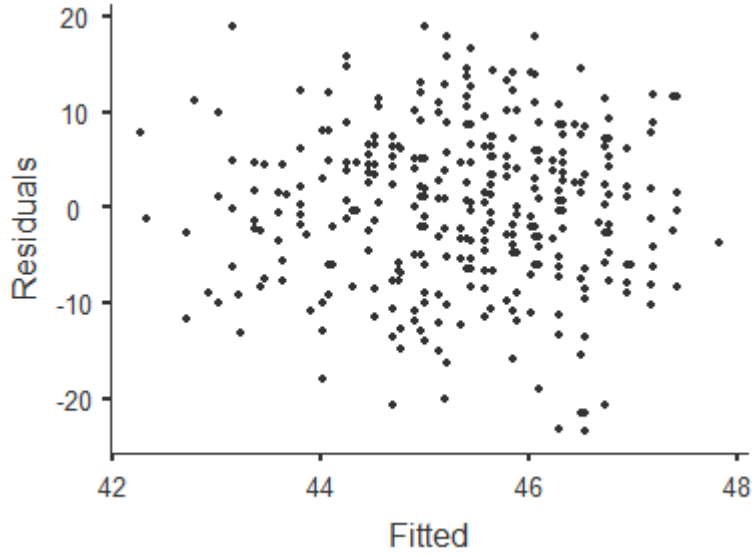


Figure M2

Scatterplot of the predicted residuals of step 2 of the TEQ-RMET-R Regression

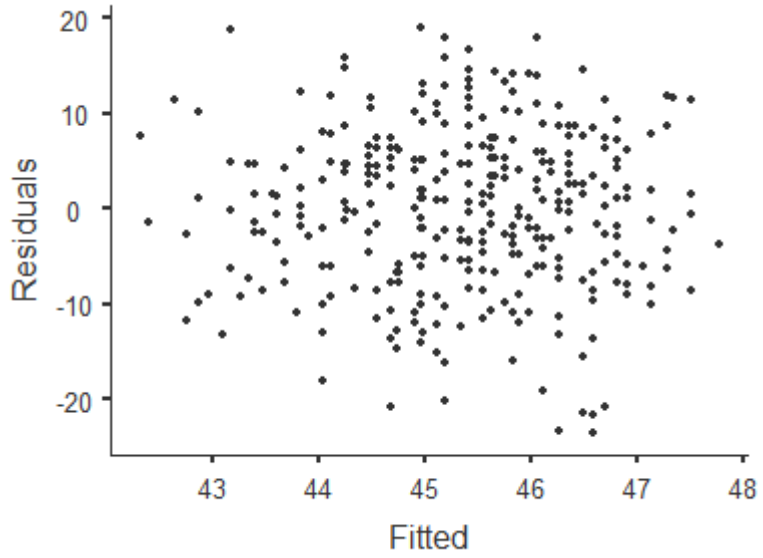


Figure M3

Scatterplot of the predicted residuals of step 1 of the TEQ-RMET-R Regression

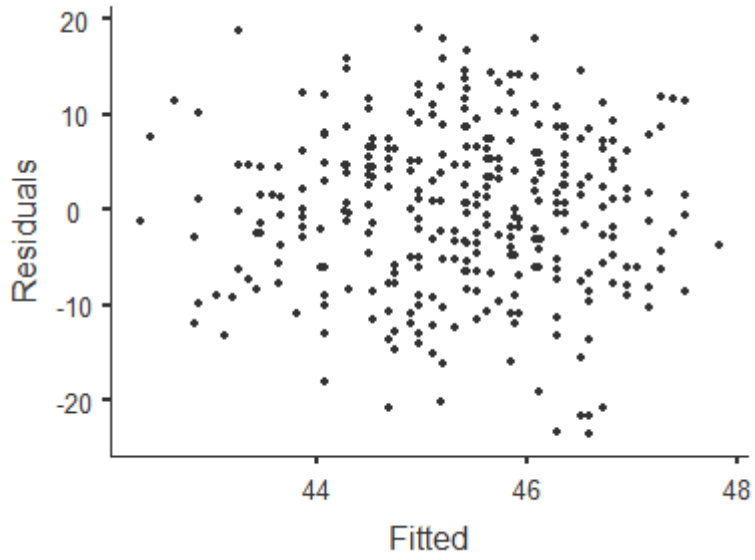


Figure M4

Scatterplot of the predicted residuals of step 1 of the MentS-RMET-R Regression

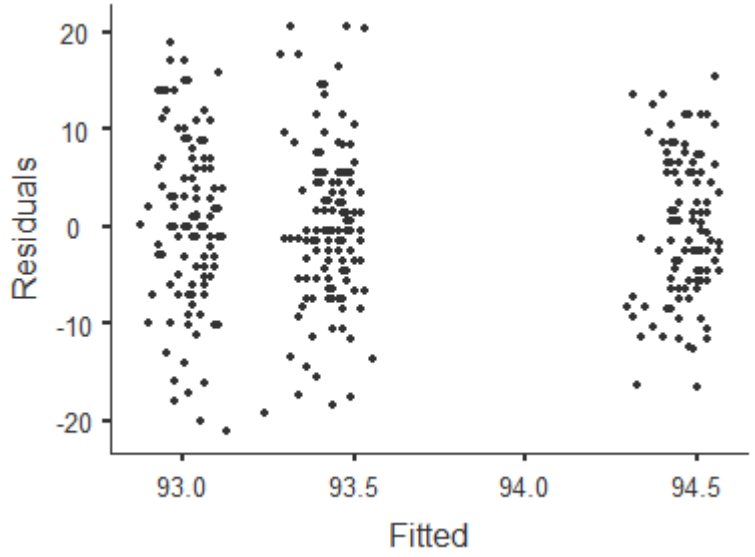


Figure M5

Scatterplot of the predicted residuals of step 2 of the MentS-RMET-R Regression

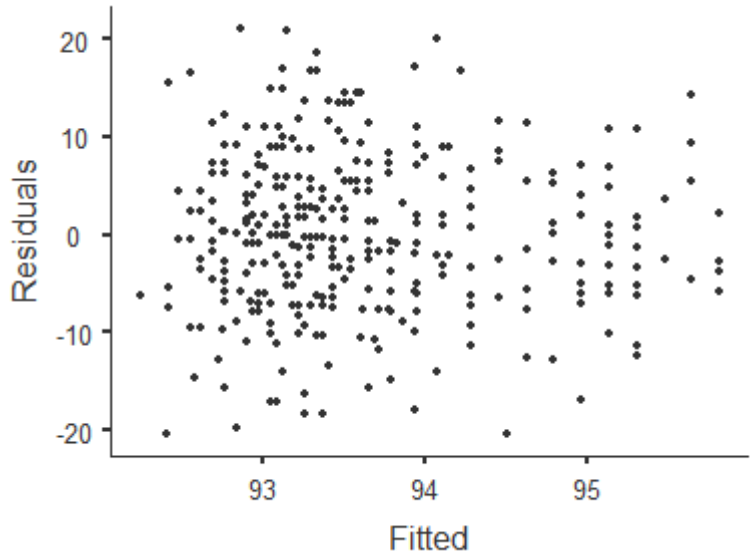
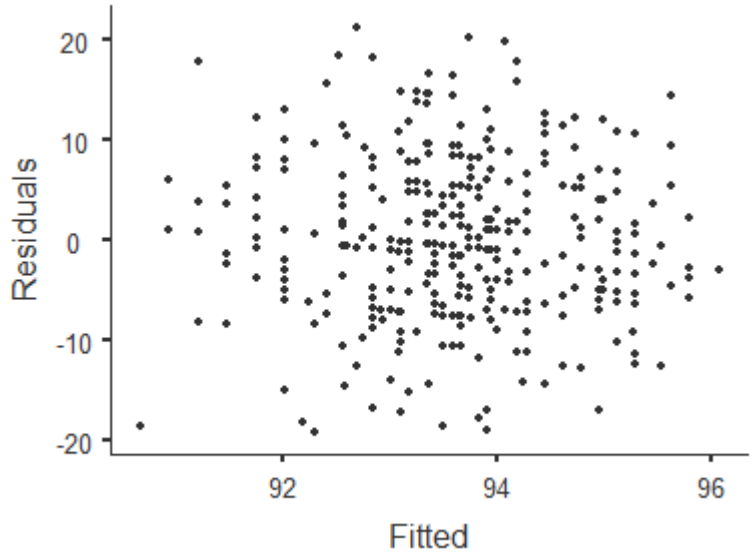


Figure M6

Scatterplot of the predicted residuals of step 3 of the MentS-RMET-R Regression



Appendix N

Q-Q plots of the standardised residuals of each regression

Figure N1

Q-Q plot of the standardized residuals of the TEQ-RMET-R regression

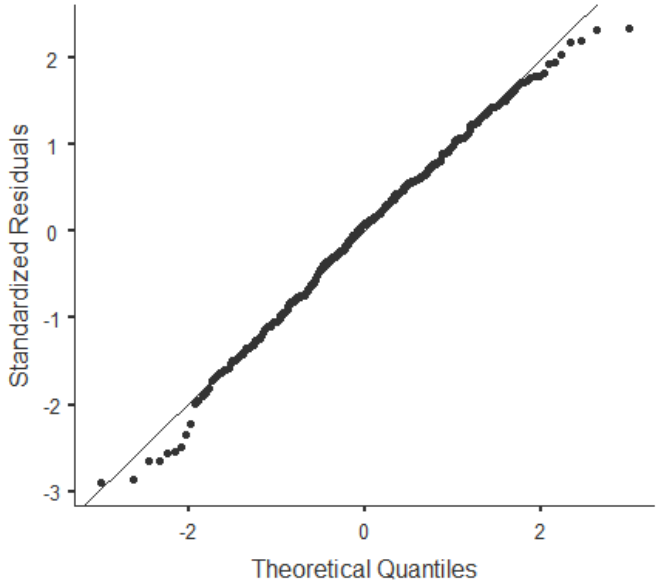
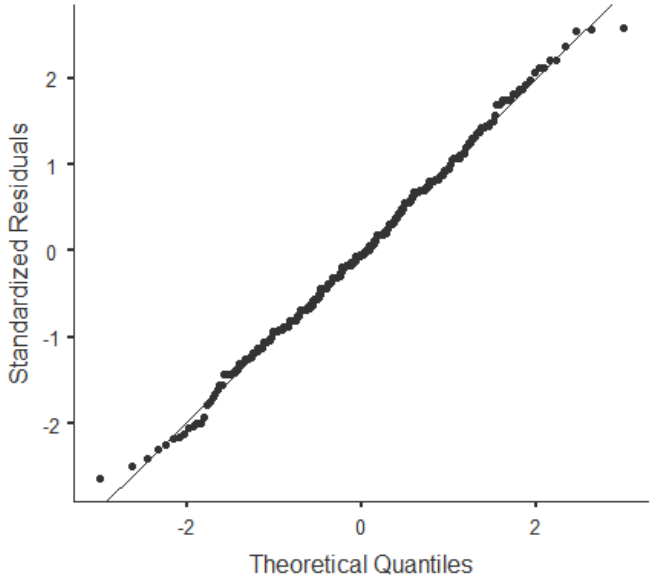


Figure N2

Q-Q plot of the standardized residuals of the MentS-RMET-R regression



Appendix O

Scatterplots of the Pearson product-moment correlation variables in this study

Figure O1

Scatterplot of Participants Age – RMET-R scores (Occluded condition)

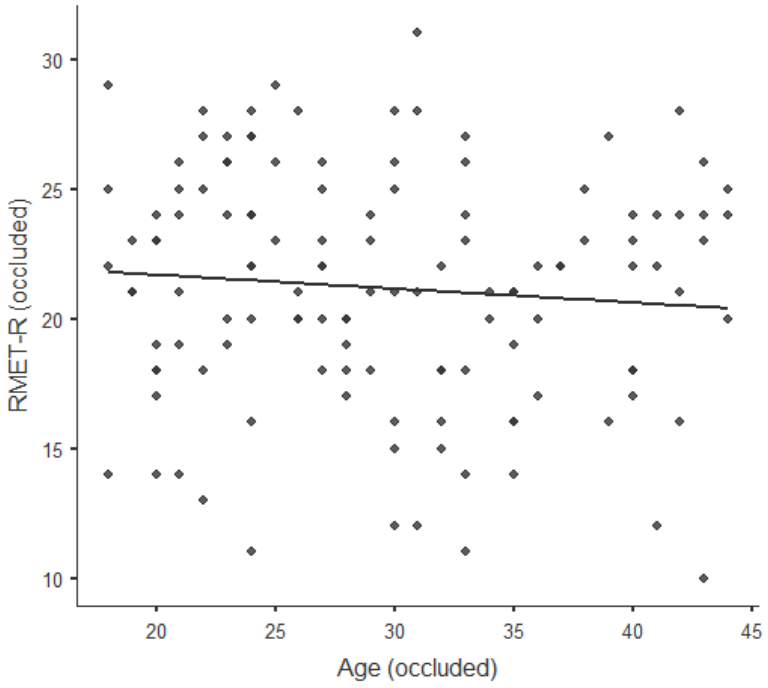


Figure O2

Scatterplot of Participants Age – RMET-R scores (Short condition)

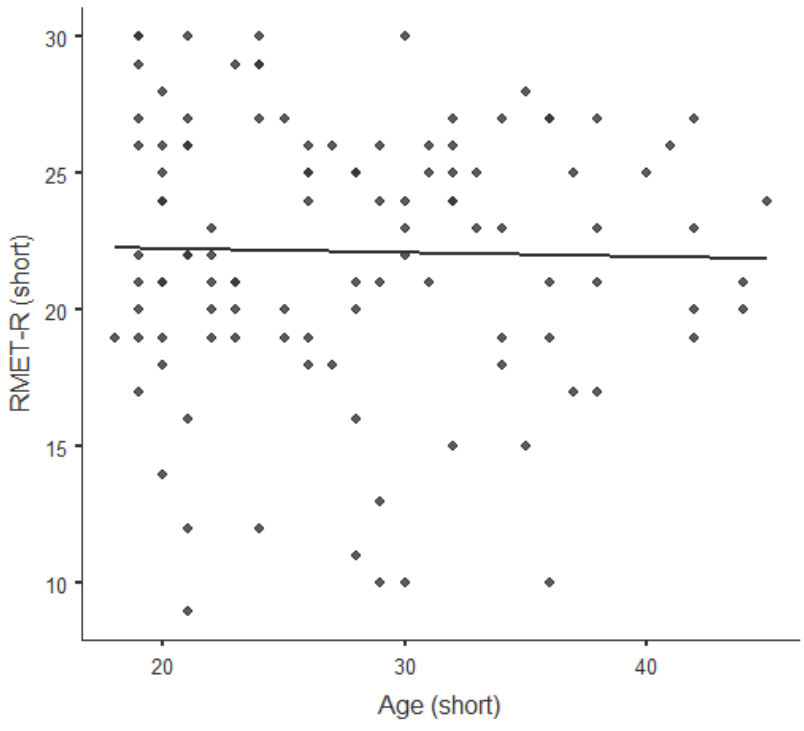


Figure O3

Scatterplot of Participants Age – RMET-R scores (Long condition)

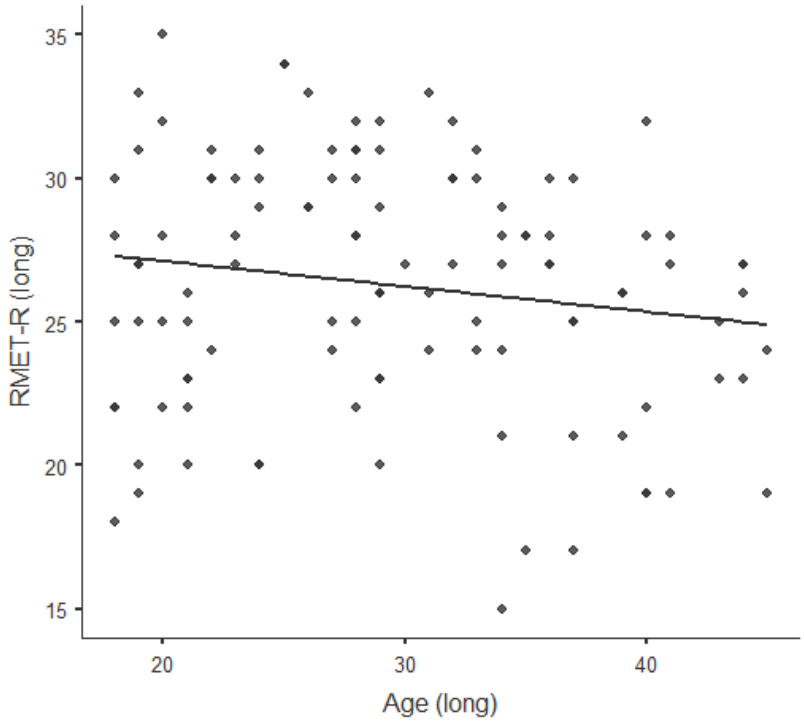


Figure O4

Scatterplot of Participants level of education – RMET-R scores (Occluded condition)

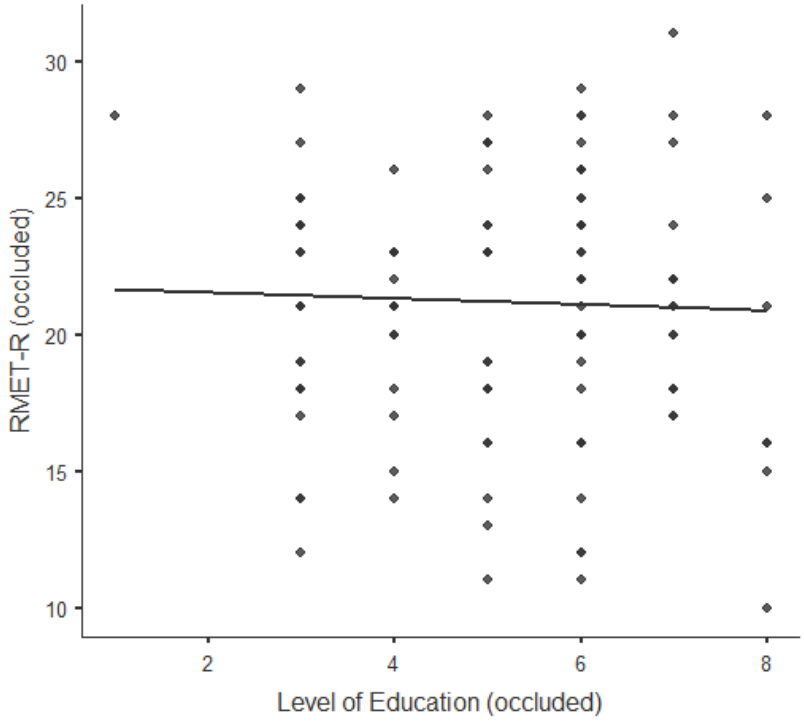


Figure O5

Scatterplot of Participants level of education – RMET-R scores (Short condition)

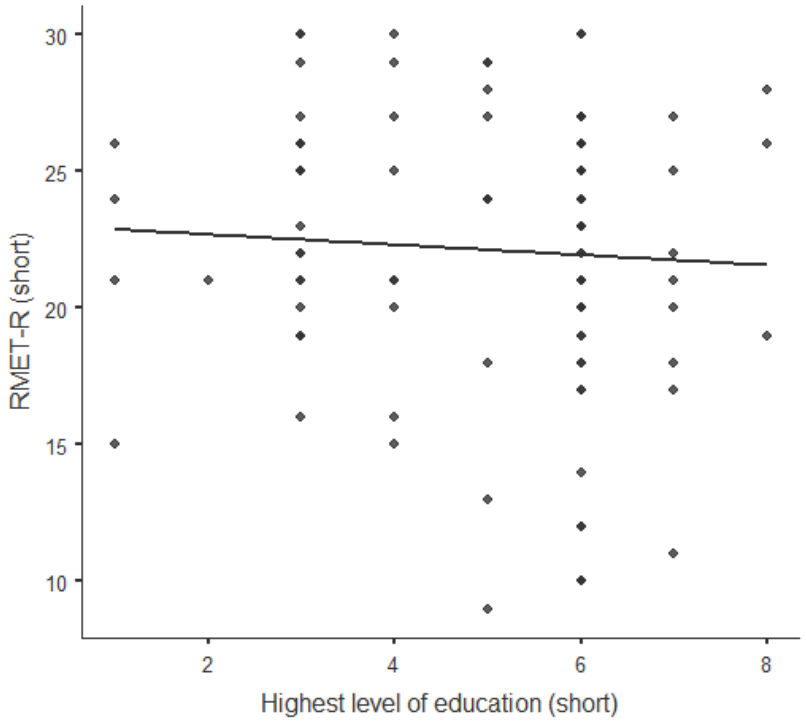


Figure O6

Scatterplot of Participants level of education – RMET-R scores (Long condition)

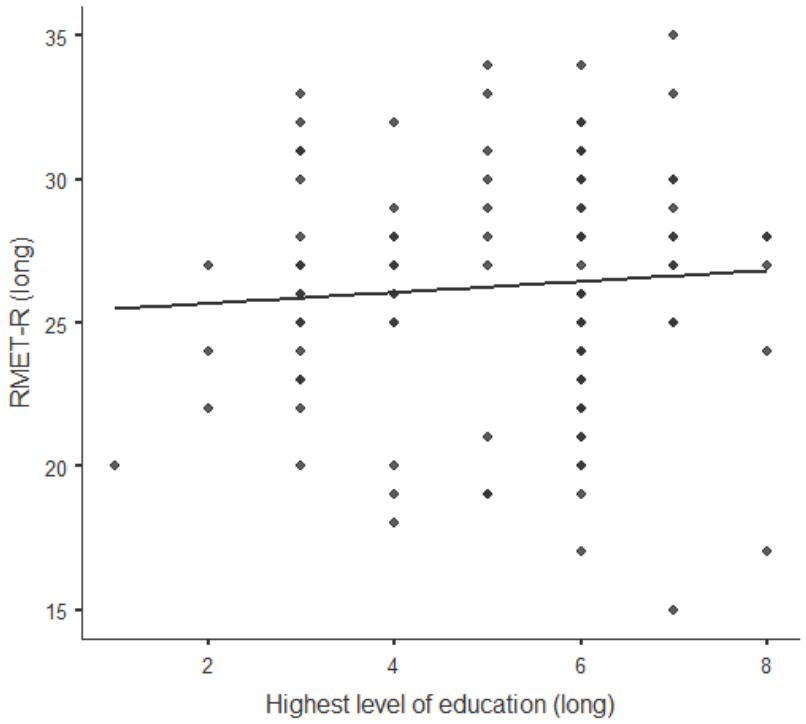


Figure O7

Scatterplot of TEQ scores – MentS scores

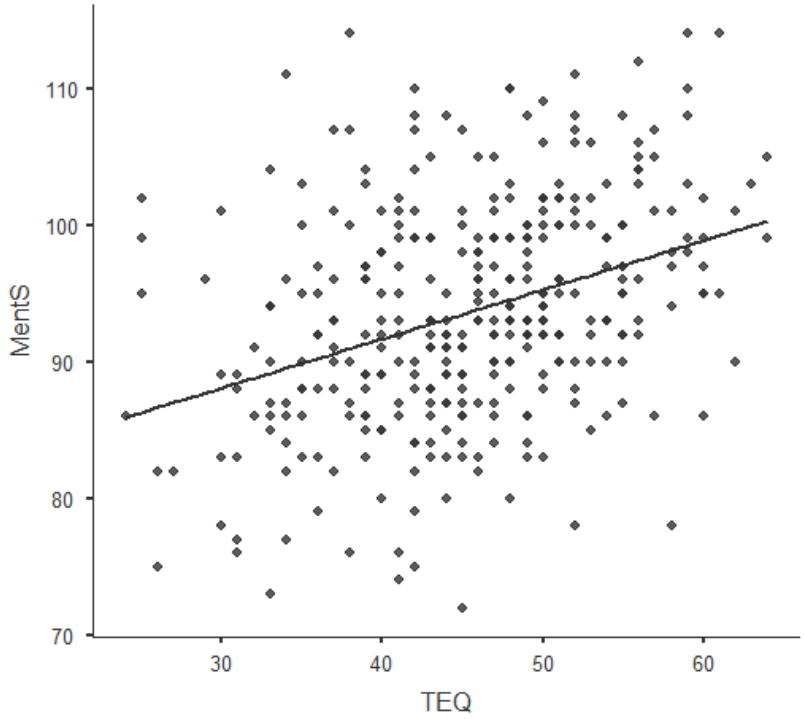


Figure O8

Scatterplot of TEQ scores- MentS Motivation scores

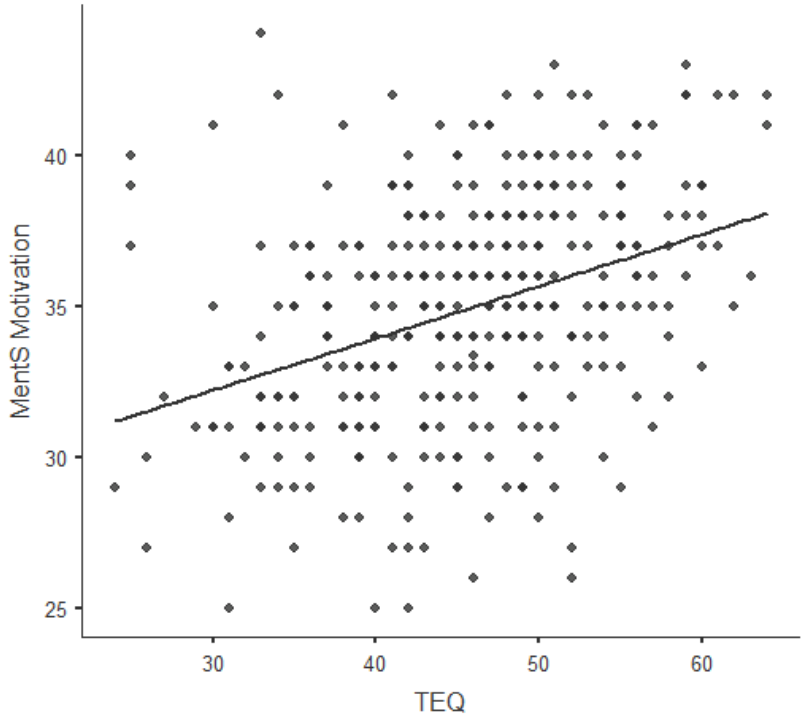


Figure O9

Scatterplot of TEQ scores- MentS Other scores

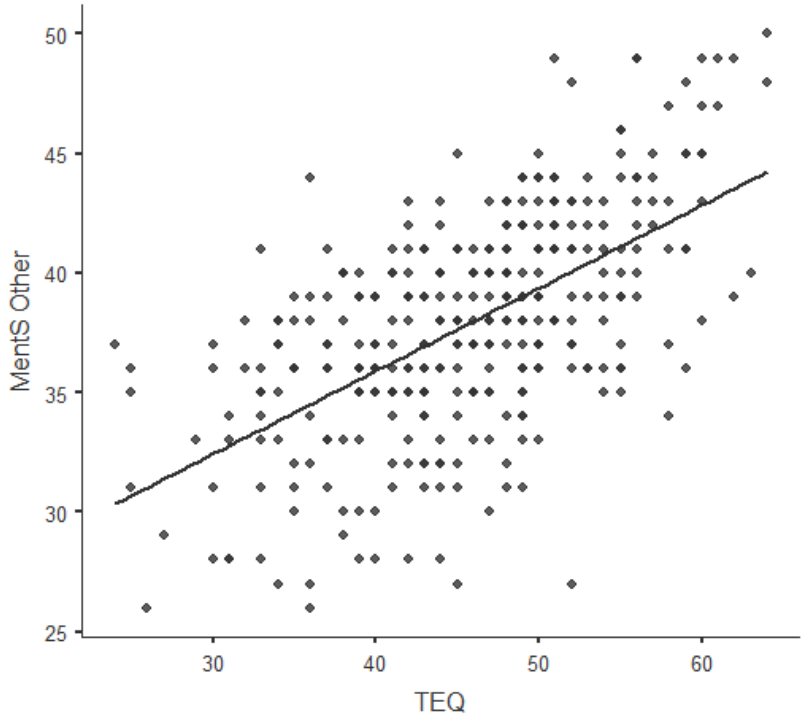


Figure O10

Scatterplot of TEQ scores- MentS Self scores

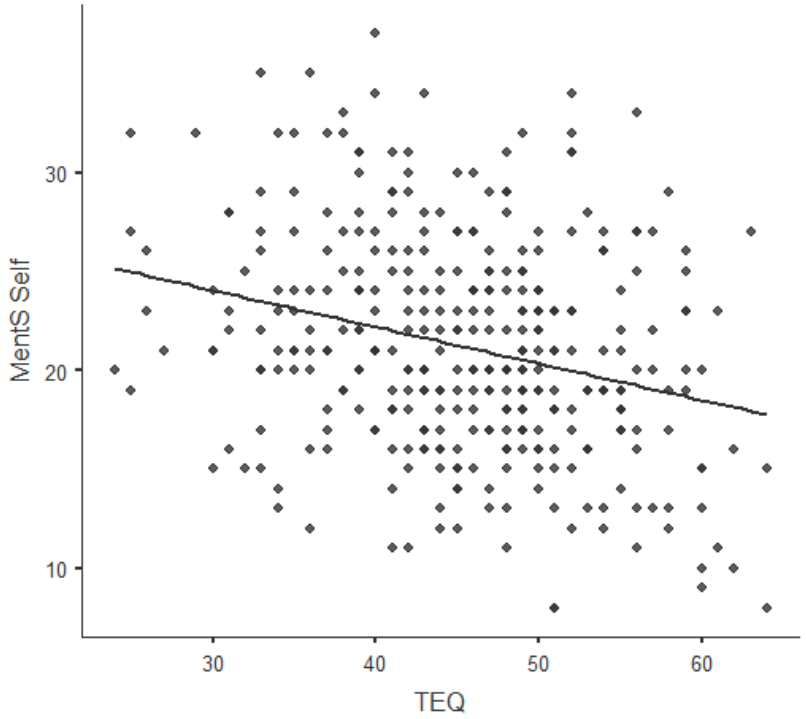


Figure O11

Scatterplot of MentS scores- MentS Motivation scores

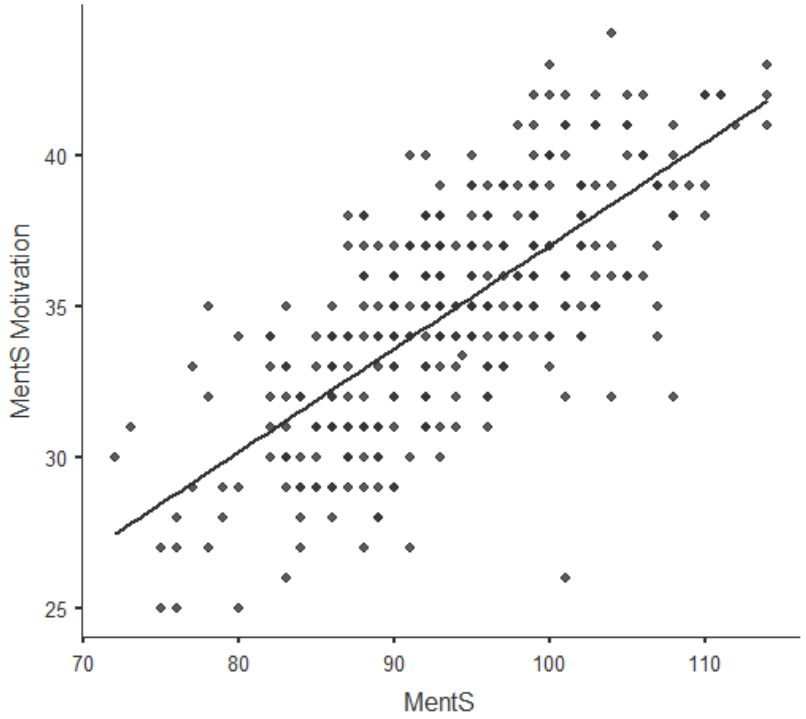


Figure O12

Scatterplot of MentS scores- MentS Other scores

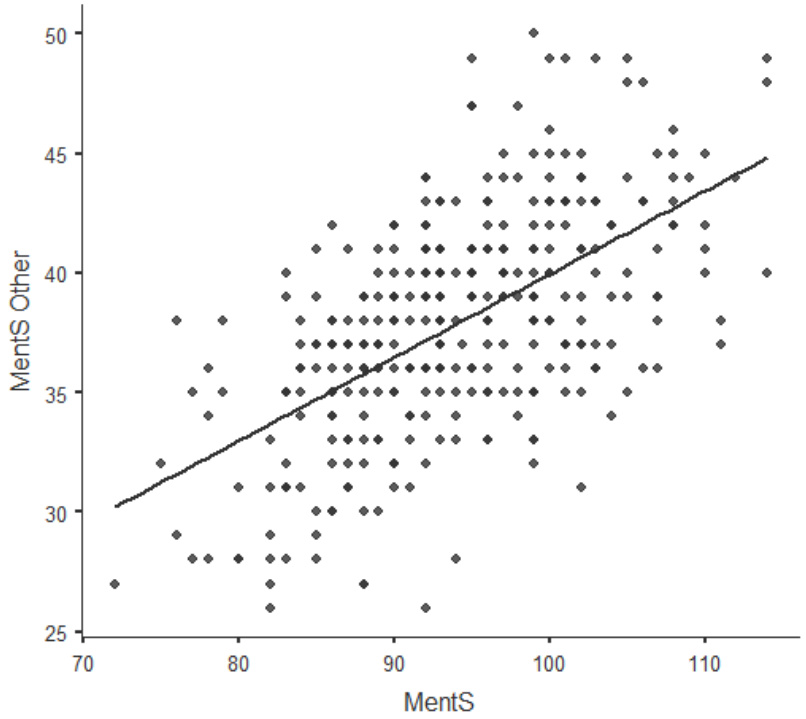


Figure O13

Scatterplot of MentS scores- MentS Self

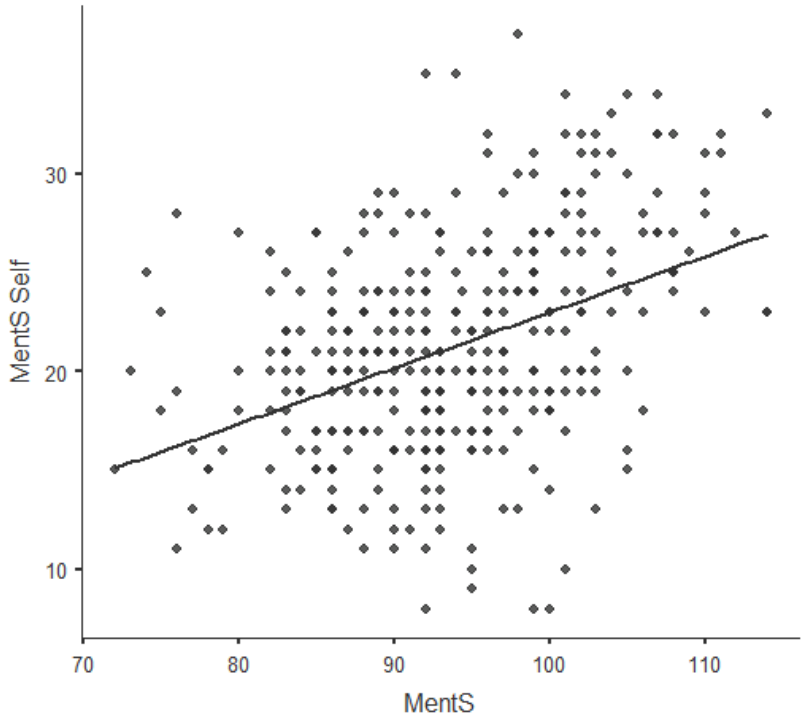


Figure O14

Scatterplot of MentS Motivation scores- MentS Other scores

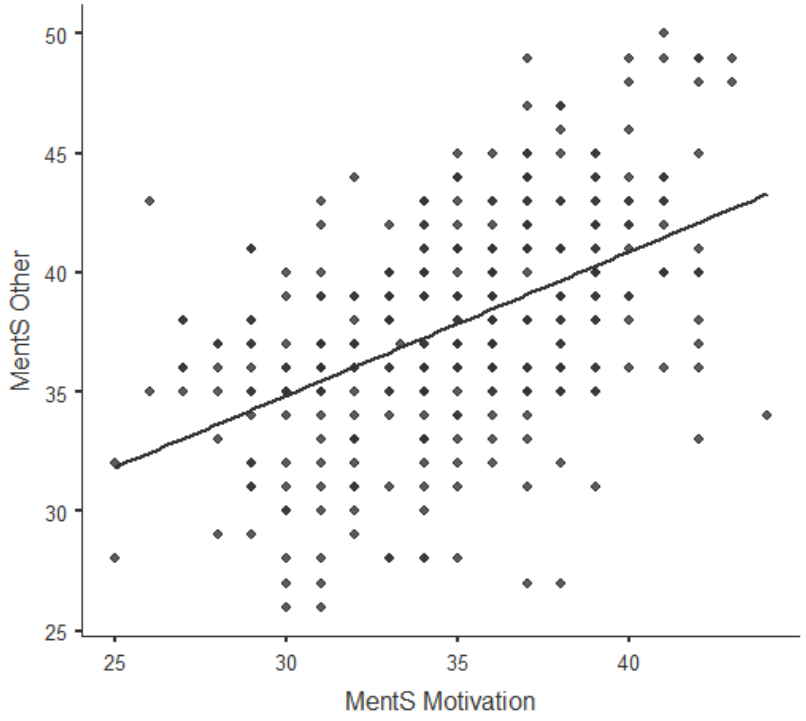


Figure O15

Scatterplot of MentS Motivation scores- MentS Self scores

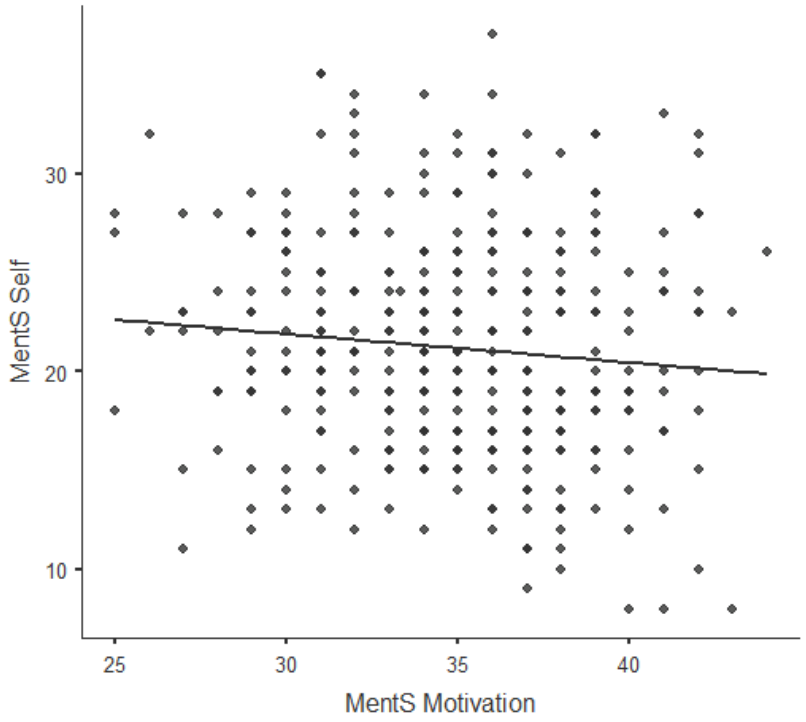


Figure O16

Scatterplot of MentS Other scores- MentS Self scores

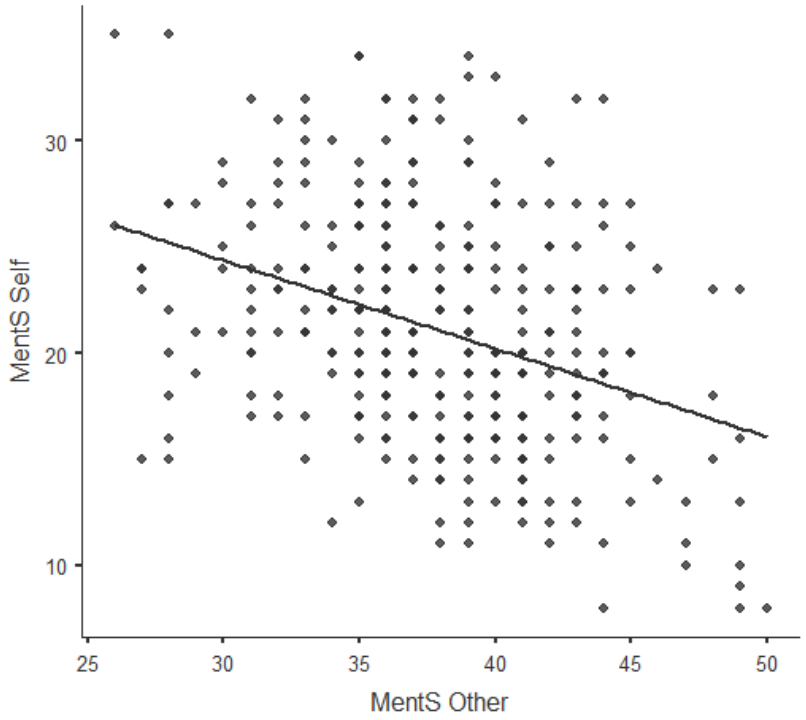


Figure O17

Scatterplot of MentS Other scores- RMET-R scores