An Empirical Evaluation of the Reliability and Validity of the "Reading the Mind in the Eyes" Test

By

Wendy C. Higgins

BA English Literature
BS Cognitive and Brain Sciences
GradCert Mathematics
HDip Linguistics
MA English Literature
MA Teaching (Secondary)

A THESIS SUBMITTED TO MACQUARIE UNIVERSITY
FOR THE DEGREE OF MASTER OF RESEARCH
DEPARTMENT OF COGNITIVE SCIENCE

Principal Supervisor: Dr Vince Polito
Associate Supervisor: Associate Professor Robyn Langdon
Associate Supervisor: Dr Robert Ross

21 December 2020

Table of Contents

List of Tables

List of Figures

Abstract

The Reading the Mind in the Eyes test (RMET) is a widely used measure of theory of mind (ToM) ability that was originally designed to detect ToM deficits in autistic adults and validated based on the performance of autistic individuals. Despite its popularity, there are questions regarding the test's factor structure, whether it taps mental state reasoning components of ToM or simply emotion recognition ability, and its validity for use in non-autistic populations. In the current study, a US representative sample of 1,181 adults completed the RMET, the Toronto Alexithymia Scale, and the Autism Spectrum Quotient. Exploratory factor analysis (EFA) on the full sample and separate EFA on individuals with high and low levels of autistic traits provided evidence for a three-factor model and two overlapping, but distinct, three-factor models for individuals with high versus low levels of autistic traits. However, the RMET had poor psychometric properties for all three groups. Hierarchical regression analysis and structural equation modelling suggested that levels of alexithymia traits and autistic traits each predict performance on the RMET. I conclude that the lack of strong psychometric properties for the RMET, evidence of variation in performance across samples, and the absence of theoretical explanations for how the test captures ToM ability undermine the validity of the RMET. I argue that until these issues are satisfactorily addressed, researchers should not use the RMET as a measure of social cognition.

Statement of Originality

This work has not previously been submitted for a degree or diploma in any university. To the best of my knowledge and belief, the thesis contains no material previously published or written by another person except where due reference is made in the thesis itself.

(Signed)_____ Date: _____
Wendy C. Higgins

Ethics Approval Reference Number: 52020625515320

Acknowledgments

I would like first to express my gratitude to my supervisors, Dr Vince Polito, Associate Professor Robyn Langdon, and Dr Robert Ross for their support with this project. You brought a diverse wealth of knowledge and experience, which enriched my own experience and the resulting project immeasurably. Thank you, Vince, for your receptive response to my ever shifting grand ideas, while keeping me on track. Thank you, Robyn, for remaining optimistic and helping me to carve out a feasible project plan when I began to fear all hope was lost. Thank you, Rob, for pointing me in the direction of useful resources and challenging me to comprehensively support all my claims.

I would also like to thank Andrew Roberts for generously sharing his time and knowledge with me as I wrapped my head around statistical analyses.

Thank you, Samuel Jones and Spencer Arbige, for sharing the MRes journey with me.

To my little lady, Isabel Higgins, and my little man, Leo Higgins, thank you for your (often gracious) willingness to support me through a demanding year. This has been a year like no other on all fronts, and I appreciate your ongoing encouragement throughout.

And lastly, I would like to acknowledge the support of my husband, Eoin Higgins. This project is a culmination of many years of study across diverse disciplines, and you have supported me through them all. Thank you for always knowing that I can do it and helping with the logistics of creating the physical and mental spaces in which to get it done.

1. Introduction

The Reading the Mind in the Eyes test (RMET, Baron-Cohen et al., 1997; Baron-Cohen, Wheelwright, Hill et al., 2001) is a measure of theory of mind (ToM) ability that was originally designed to detect subtle ToM deficits in autistic adults. Since then, the RMET has been used as a measure of social cognition in a range of additional clinical populations, including individuals with anorexia nervosa (Russell, et al., 2009), schizophrenia (Li et al., 2020), social anxiety disorder (Washburn, et al., 2016), depression (Harkness, et al., 2010), and bipolar disorder (Bora et al., 2016). The RMET has also been used as a measure of individual differences in ToM ability within nonclinical populations (Black, 2019). However, as argued in this paper, research reporting the psychometric properties of the RMET has yielded mixed results. Although it is widely used, questions related to the reliability and validity of the RMET suggest that performance may vary across populations and it may not be fit for purpose.

In this paper, I evaluated the reliability and validity of the RMET in a sample that is both larger than previous validation studies (e.g. Preti et al., 2017; Prevost et al., 2014; Vellante et al., 2013) and demographically representative of the US population. Noting that the RMET was originally validated through the performance of autistic individuals (Baron-Cohen, Wheelwright, Hill et al., 2001), to address inconsistent findings in previous factor analyses of the RMET, I tested the possibility that the factor structure of the RMET is different for individuals with higher versus lower levels of autistic traits. To evaluate the validity of the RMET as a measure of ToM, I explored the relationship between scores on the RMET and a battery of alternative tasks related to social cognition, including measures of alexithymia, autistic traits, comfort viewing eye stimuli, and an alternative ToM task. As a secondary set of analyses, I used confirmatory factor analysis (CFA) to evaluate the proposed factor structures of the self-report measures of alexithymia and autistic traits when delivered online to a large, demographically representative US sample.

1.1 What is Theory of Mind?

ToM is a component of social cognition that underlies the ability to infer what other people are thinking and feeling. The term ToM was first introduced by Premack and Woodruff (1978, p. 515), who were exploring whether chimpanzees have a capacity similar to humans whereby they can impute mental states such as knowledge, belief, and purpose to understand, reason about, and predict the behaviour of others. Premack and Woodruff succinctly defined ToM as the ability to "impute mental states to oneself and to others." This definition is still widely accepted as the overarching definition of ToM; however, various subtypes have been proposed in the literature. Some researchers divide ToM into subtypes based on the type of mental state being inferred, drawing a distinction between reasoning about affective mental states, which have an emotional component (affective ToM), and reasoning about purely cognitive mental states (cognitive ToM, Shamay-Tsoory & Aharon-Peretz, 2007). Other researchers categorise ToM as a two component process, proposing a distinction between social-perceptual ToM, which involves decoding or identifying mental states, and social-cognitive ToM, which involves reasoning about mental states (Mısır et al., 2018).

There is a lack of clarity in the literature on the boundaries between ToM, emotion recognition, and empathy. While emotion recognition and ToM are both aspects of social cognition, some researchers consider emotion recognition to be a component of ToM while others consider it to be a separate cognitive capacity (Marsh et al., 2016; Oakley et al., 2016; Russell et al., 2009). The term 'cognitive empathy', which refers to a subtype of empathy, also overlaps with ToM to the extent that some authors (Lawrence et al., 2004; Richard-Mornas et al., 2014) use the terms interchangeably.

Ongoing debates about how to define ToM are relevant to assessing to validity of the RMET as a measure of ToM. For example, if the RMET turns out only to measure emotion recognition ability, then its status as a measure of ToM will depend on whether a researcher classifies emotion recognition as a component of ToM or a separate cognitive ability. This issue is not unique to the

RMET. As noted by Schaafsma et al. (2015), the current variety of conceptions of ToM and the diversity of tools used to measure it limit our ability to cohesively further our understanding of this facet of social cognition and we need to build a tractable definition of ToM for use in empirical research studies. In the meantime, researchers need to be explicit about what they mean by the term ToM and how they are operationalising it. For the purposes of this study, I acknowledge that ToM is a multifaceted component of social cognition. However, in the absence of an agreed upon cognitive model of ToM, I remain agnostic as to the specific social cognitive processes that ToM comprises, while acknowledging how different definitions potentially impact the validity of the RMET.

1.2. The Development of the RMET

Baron-Cohen (1995) proposed that autism results from a deficit in ToM ability. Noting that some autistic adults could pass existing ToM tasks such as second-order false belief tasks, Baron-Cohen et al. (1997) suggested that ToM deficits in autistic adults might be too subtle for existing ToM tasks to measure. The RMET was created to identify ToM deficits in autistic adults that were not well captured by other measures (Baron-Cohen et al., 1997, Baron-Cohen, Wheelwright, Hill et al., 2001).

According to Baron-Cohen (1995), humans possess a ToM mechanism that relies heavily on information derived from people's eyes, including expressions and gaze direction. The RMET was designed to assess the ability to use information from people's eyes to infer mental states. The stimuli consist of black and white images of people's eyes that were collected from magazines and a selection of mental state terms. The original version of the task paired two mental state terms with opposite meanings to each image (Baron-Cohen et al., 1997), but the test was revised to include four mental state terms of similar valence (see Figure 1) because the original version was too easy (Baron-Cohen, Wheelwright, Hill et al., 2001). Most researchers use the revised version (e.g. Adams et al., 2010; Fertuck et al., 2009).

Figure 1

RMET Item 34

Note. Image retrieved from the Autism Research Centre (2020) https://www.autismresearchcentre.com/tests/eyes-test-adult/

Because the images were collected from magazines rather than created specifically for the task, there are no objectively correct responses. The targets were created by the test's authors and validated in a sample of 225 participants consisting of members of the general public and students from Cambridge University. The criteria for a test item to be validated were that at least 50% of participants chose the target and no more than 25% selected the same foil. Thirty-six out of forty items met these criteria and make up the current revised version of the RMET.

The RMET is currently one of the most frequently used measures of ToM (Eddy, 2019). Despite its widespread use, the evidence for the reliability and validity of the RMET is both limited and inconclusive, as discussed below.

1.3. Psychometric Properties of the RMET

1.3.1. Reliability

Where reported, the test-retest reliability of the RMET is generally acceptable, whereas levels of internal consistency are highly variable. There are no strict cut off values for Cronbach's alpha to indicate acceptable levels of internal consistency, however, minimum values of .70 or .75 are often cited (Christmann & Van Aelst, 2006). As shown in Table 1, reported alpha levels for the RMET often fall below this range. Moreover, Cronbach's alpha can be artificially inflated for longer tests (Tavakol & Dennick, 2011). For example, Zinbarg et al. (2006) found an increase from twelve to twenty items impacted alpha values. McDonald's omega has been recommended as an alternative measure of internal consistency (Flora, 2020). While there is no set cutoff value for omega (Green & Yeng, 2015), some researchers also use the value of ≥ .70 as indicative of acceptable reliability (Bado et al., 2018).

One potential explanation for the low levels of internal consistency of the RMET is that the test has a multifactorial structure. Olderbak et al. (2015) evaluated this possibility using EFA on the RMET scores of 484 participants collected online via Amazon Mechanical Turk. EFA indicated a five-factor model, which the authors rejected as the factors had no obvious conceptual interpretation, and even with five factors, nine items (25%) failed to load on to any of the factors.

Conducting EFA on the German child version of the RMET in a sample of 596 seventh to ninth graders, Müller and Gmünder (2014) also failed to identify a satisfactory factor structure. Their first two factors only accounted for a small amount of variance (13.7%), and only one item had a factor loading above 0.5.

Table 1

Reported Internal Consistency and Test-Retest Reliability for the RMET

Study	Language	Cronbach's Alpha	Omega	Intraclass Correlation
Black (2019)	English	.78		
Burke et al. (2016)	English	.73		
Charernboon & Lerthattasilp (2017)	Thai	.70		.92
Espinós et al. (2018)	Spanish	.71	.78	
Fossati et al. (2017)	Italian	.64		
Giordano et al. (2019)	Spanish	.53		
Girli (2014)	Turkish	.71		
Jankowiak-Suida et al. (2016)	Polish	.67		.89
Khorashad et al. (2015)	Persian	.37		.74
Kotrla Topić & Perković Kovačević (2019)	Croatian	.54	.74	
Kung (2020)	English	.63		
Mar et al. (2006)	English	.60		
Meyer & Shean (2006)	English	.48		
Oakley et al. (2016)	English			
Olderbak et al. (2015)	English		.75	
Öztürk et al. (2020)	Turkish	.84		
Prevost et al. (2014)†	French	.53	.70	
	English	.77	.70	
Sadeghi Bahmani et al. (2018)	Persian	.79		
Schmitt et al. (2020)††	Mandarin		.68	
	German		.69	
Vellante et al. (2013)	Italian	.61		.83

Note. These studies were collated from a Google Scholar search for "reading the mind in the eyes" AND alpha OR omega" and are the first 20 studies that included an alpha or omega value for the RMET. †Provost et al. (2020) had both an English and French sample. They only reported a single test-retest reliability figure. ††Schmitt et al. (2020) compared versions of the RMET using both white and Asian eyes in German and Chinese samples. Reported omega values are for the white eyes.

Olderbak et al. (2015) evaluated a conceptually driven three-factor model proposed by Harkness et al. (2005), in which test items were divided into positive, negative, and neutral factors based on the valence of each item. However, they ultimately rejected the three-factor valence model. They found satisfactory model fit according to root mean square error of approximation (RMSEA, .019, see sections 7.4. and 7.4.1. for a description of model fit statistics), but model fit was poor according to comparative fit index (CFI, .75) and the factor loadings were weak. Vellante et al. (2013) also rejected the three-factor valence model for the Italian version of the RMET in a sample of 200 university students. They reported that a three-factor model failed to converge. When they forced a three-factor model, they found acceptable model fit according to the standardised root mean square of residuals (SRMR, .072), but the CFI was very low (.310), and items did not consistently load on to the anticipated factors. As noted by Hudson et al. (2020), one issue with the valence factor model is that different researchers have categorised items differently. It is difficult to conclusively categorise test items as belonging to specific valence categories, as a mismatch between the valence of the image and the target term has been found for some RMET test items (Scott et al., 2011).

In contrast, the authors of a recent study of the Korean version of the RMET that uses images of Asian eyes claimed good fit for the three-factor valence based model based on data from 200 adults from the general population. However, while model fit was good according to RMSEA (.04), as with the other studies, their reported CFI (.450) indicated poor model fit.

Another interpretation of the low internal consistency of the RMET is to assume that the RMET has a unitary factor structure and that the poor internal consistency results from the inclusion of inappropriate test items. In this case, an appropriate strategy is to identify the problematic test items and remove them in order to improve model fit and internal consistency. Olderbak et al. (2015) took this approach. CFA resulted in poor model fit for a single factor. To address this, they proposed a reduced 10-item, single factor version of the RMET that had acceptable psychometric properties. Harkness et al. (2010) also proposed a shortened version of the RMET in response to low

internal consistency in their data (Cronbach's α = .58). They removed eight problematic RMET test items to create a 28-item version of the test, which increased Cronbach's alpha to .63. Harkness et al. considered this value sufficient for experimental use, however, it is still below the recommended level of > .70. Using item response theory, Black (2019) found that a single factor solution with a reduced 20-item test produced the best model fit (RMSEA = .17, TLI = .971). Factor analyses of the Serbian (Dordevic et al., 2017) and Spanish (Redondo & Herrero-Fernandez, 2018) versions of the RMET also resulted in proposals for shortened versions that remove approximately half of the original test items due to low internal consistency and poor model fit for the full 36-item test.

Concerningly, there is more variation than consistency in the test items that constitute these short form versions of the RMET. For example, there is only one item retained (item 8) and four items excluded (items 3, 6, 21, 25) across all versions of the RMET reviewed in Table 2. The authors of these abbreviated versions did not provide theoretical explanations for why the rejected test items might be less valid measures of ToM, or conversely, what properties the retained items have that enable them to validly capture ToM ability. In summary, previous research has shown a significant level of variability in the reliability of the RMET across different samples.

1.3.2. Validity

The RMET aims to assess ToM capacity via a person's ability to select a mental state term that matches an image of a person's eyes. When validating the original version of the RMET (Baron-Cohen et al., 1997) with two mental state terms with opposite meanings per image (e.g. concerned vs. unconcerned), Baron-Cohen and colleagues reported evidence of convergent validity with another ToM measure, the Strange Stories task which requires participants to answer questions about brief stories that include ToM related content including jokes, white lies, persuasion, and misunderstanding (Happé, 1994). They also reported that autistic individuals performed worse on the RMET than individuals without autism. In conjunction with Baron-Cohen's (1995) theory that autism results from underlying deficits in ToM ability, they claim that the lower scores provide additional support for the validity of the RMET as a measure of ToM ability.

Table 2

Reduced Versions of the RMET

RMET item		Target	Black (2019)	Olderbak et al. (2015)	Harkness et al. (2010)†	Redondo & Herrero-Fernandez (2018)	Dordevic et al. (2017)
			English N = 591	English N = 484	English N = 93	Spanish N = 433	Serbian N = 260
1		playful			✓		
2		upset			✓	✓	
3		desire					
4		insisting	✓		✓		✓
5		worried	✓		✓	✓	
6		fantasizing					
7		uneasy	✓	✓	✓		
8		despondent	✓	✓	✓	✓	✓
9		preoccupied	✓		✓		✓
10		cautious	✓			✓	✓
11		regretful	✓		✓	✓	✓
12		sceptical		✓	✓	✓	✓
13		anticipating			✓	✓	✓
14		accusing		✓	✓	✓	✓
15		contemplative	✓	✓	✓	✓	✓
16		thoughtful	✓		✓		✓
17		doubtful			✓	✓	✓
18		decisive	✓		✓	✓	✓
19		tentative	✓	✓	✓		✓
20		friendly			✓	✓	
21		fantasizing					
22		preoccupied		✓	✓	✓	
23		defiant			✓	✓	
24		pensive	✓	✓	✓		✓
25		panicked					
26		hostile			✓	✓	
27		cautious	✓		✓		✓
28		interested	✓		✓		✓
29		reflective				✓	
30		flirtatious	✓		✓		✓
31		confident				✓	
32		serious	✓	✓	✓		
33		concerned	✓		✓		
34		distrustful	✓		✓	✓	✓
35		nervous	✓		✓	✓	
36		suspicious	✓	✓	✓	✓	

Note. Ticks represent items that were retained in each study. Only one item was retained across all studies (item 8) and only four items were excluded in all studies (items 3, 6, 21, 25). †Harkness et al. (2010) removed eight items, but only reported seven of the removed items.

Baron-Cohen, Wheelwright, Hill et al. (2001) revised the RMET by increasing the number of mental state choices per test item from two to four because the original task did not allow for enough variation in scores above chance levels and some items were too easy. Validation of the revised RMET involved a comparison between participants' RMET scores and scores on the Autism Spectrum Quotient (Baron-Cohen, Wheelwright, Skinner et al., 2001), which is a self-report measure of autistic traits. No validation with another established ToM task was demonstrated, thus the primary source of validation for the revised RMET was the correlation between levels of autistic traits and performance on the RMET and the poorer performance of autistic individuals versus a control group.

1.3.2.1. Convergent Validity.

Research on the convergence of the RMET with other ToM measures in subsequent research is limited and yields mixed results. In a sample of 100 autistic adolescents, Jones et al. (2018) found a correlation between performance on the children's RMET and three other ToM tasks, Happé Strange Stories task ($r = .29$, $p < .01$), a false belief task adapted from Sullivan et al. (1994, $r = .45$, $p < .001$), that requires an understanding that others can have beliefs that are untrue, and Frith-Happé animation task ($r = .45$, $p < .001$, Abell et al., 2000), in which participants watch short videos of shapes moving either randomly or in ways that can be interpreted as social interactions and describe what they see. Ferguson and Austin (2010) also found a correlation between RMET scores and performance on the faux pas task ($r = .28$, $p < .01$, Gregory et al. 2002; Stone et al. 1998) in a sample of 162 participants consisting of university students and members of the general public. The faux pas task involves hearing a story and indicating whether anyone has said anything socially inappropriate, and it is considered to be a measure of both cognitive and affective ToM. However, Ahmed and

Miller (2011) failed to find a correlation between the RMET and the Strange Stories task or the faux pas task in a sample of 135 university students. Looking at the performance of both children and adults on a range of ToM tasks, Warnell and Redcay (2019) failed to find a correlation between the children's versions of the RMET and the faux pas or Strange Stories task. They also failed to find a correlation between performance on the adult RMET and performance on ToM tasks including a story task and a task evaluating pragmatic language ability. The lack of convergence may relate to the variety of different cognitive abilities targeted by these tasks, which again emphasises the challenge that varied definitions of ToM pose to empirical ToM research.

Despite unclear evidence on the convergent validity of the RMET with other established ToM tasks, this test is very frequently used (2,876 citations on Web of Science as of December 16, 2020), and it appears that many researchers may be unaware of its psychometric shortcomings. In some cases, researchers have even stated that the RMET has convergent validity with other ToM tasks but have only referenced the original paper introducing the revised RMET, which did not assess convergent validity (e.g. Franklin & Zebrowitz, 2016) or failed to provide any supporting references (e.g., Adams et al., 2010).

In addition to the limited evidence of convergence between the RMET and other ToM tasks, two intertwined issues complicate evaluations of the validity of the RMET as a measure of ToM: variability in how researchers define ToM and variability in the ways in which researchers use the RMET in their research. While the RMET was originally designed as a measure of ToM ability in adults, it is also described in the empirical literature as a measure of affective ToM (Raffo De Ferrari et al., 2015; Rominger et al., 2016), social-perceptual ToM (Ferguson & Austin, 2010) cognitive empathy (Warrier et al., 2018; Youssef et al., 2014), and emotion recognition (Harrison et al., 2010; Pahnke et al., 2020; Vellante et al., 2013). As noted above, 'cognitive empathy' has been used as a direct synonym of ToM by some authors (Lawrence et al., 2004; Richard-Mornas et al., 2014), however other researchers treat these as distinct abilities and use separate tools to measure them (Dziobek et al., 2006).

The debate in the literature over the relationship between emotion recognition, empathy, and ToM also impacts questions concerning the validity of the RMET as a measure of ToM. A number of papers assess the validity of the RMET through convergence with the Empathy Quotient (Baron-Cohen et al. 2004) and the Toronto Alexithymia Scale (TAS, Bagby et al., 1994), which index empathy and emotion recognition abilities respectively (Lee et al., 2018; Vellante et al., 2013). Conversely, when validating another ToM task, Brewer et al. (2017) specifically looked for the absence of a correlation with self-reported empathy in order to validate their task as a measure of ToM ability. The validity of the RMET as a measure of ToM will vary according to how researchers define ToM.

1.3.2.2. Construct Validity.

Research on the construct validity of the RMET is also limited. To my knowledge, there are no published evaluations of the theoretical bases of the RMET laid out by Baron-Cohen and colleagues. Instead, many researchers cite the ability of the RMET to distinguish between individuals with and without autism spectrum disorders as support for its use as a measure of ToM (Adams et al., 2010; Baron-Cohen et al., 2015). As noted above, the basis for this claim is the theory that a ToM deficit underlies autism (Baron-Cohen, 1995). Recently, Gernsbacher and Yergeau (2019) have challenged this claim. This highlights the need for more rigorous theoretical evaluation of the construct validity of the RMET and leaves open the possibility that the RMET relies on an ability other than ToM that autistic individuals also find challenging.

Alexithymia (a condition in which an individual has difficulty recognising and describing their own and others' emotions) frequently co-occurs with autism spectrum disorders. Oakley et al. (2016) argued that the poorer performance of autistic individuals on the RMET is better explained by co-occurring alexithymia and a deficit in emotion recognition, rather than a deficit in ToM ability. They supported this claim with a study in which they found that alexithymia is more predictive of performance on the RMET than an autism diagnosis or severity of autistic traits as measured by the Autism Spectrum Scale. The theoretical significance of this finding depends on whether emotion recognition is considered a part of ToM or a separate ability. If emotion recognition is a part of ToM

ability, then the RMET could still be categorised as a measure of ToM, however, if emotion recognition is a separate cognitive ability, then Oakley et al.'s (2016) findings suggest that the RMET is not a measure of ToM ability.

Consistent with the findings of Oakley et al. (2016), Gökçen et al. (2016) found higher levels of alexithymia were associated with poorer performance on the RMET in individuals with higher levels of autistic traits, and Tayfun and Semra (2019) found that higher levels of alexithymia were associated with poorer performance on the RMET in parents of autistic children.

1.4. Potential Variation of Factor Structure in Different Populations

One possible explanation for the variation of results seen across previous factor-analytic studies of the RMET is that different individuals use different strategies and/or combinations of cognitive capacities to complete the task or that different features of the stimuli have differential effects in different populations. For example, if individuals with high levels of autistic traits (in both clinical and nonclinical populations) use different strategies to complete the RMET than individuals with low levels of autistic traits, then this difference could potentially confound factor analyses. There is some evidence that this may be the case. In a systematic review of studies that included measures of autism, IQ, and RMET performance, Peñuelas-Calvo et al. (2019) found that, while RMET performance was correlated with verbal IQ and not performance IQ in nonclinical populations, the opposite pattern of results held for autistic individuals. They suggest that this could result from autistic individuals using different cognitive abilities to complete the RMET.

Autistic traits are believed to be normally distributed within the population (Ruzich et al., 2015). Gökçen et al. (2016) found that within a typically developing population, level of autistic traits negatively correlated with RMET scores. If autistic individuals do use different strategies to complete the RMET, it is possible that individuals without a diagnosis of autism who have high levels of autistic traits also use different strategies when completing the RMET in comparison to individuals with lower levels of autistic traits. If this is the case, a future step would be to identify the factors (e.g. gaze direction) that influence the RMET scores of individuals with high or low levels of autistic traits.

Different levels of comfort viewing eye stimuli could be a factor that differentially influences autistic individuals' RMET performance. There is evidence that in addition to having difficulty with ToM and emotion recognition tasks, autistic individuals find looking at other people's eyes stressful. Hadjikhani et al. (2017) found that autistic individuals had abnormally high levels of activation in subcortical brain regions when their gaze was restricted to the eye region of a facial stimulus. In a qualitative study, Trevisan et al. (2017) found that autistic individuals reported adverse reactions to eye contact that included threat responses, anxiety, and a sense of violation. The aversive nature of the stimuli used in the RMET might therefore predispose autistic individuals toward poorer performance, regardless of their ToM or emotion recognition ability.

The RMET stimuli contain a mixture of direct and averted gaze images. In addition to discomfort viewing eye stimuli in general, gaze direction might differentially impact autistic individuals, making some of the RMET items more aversive than others. Kylliäinen and Hietanen (2006) found that autistic children, but not controls, had a stronger skin conductance response (an indicator of arousal) to direct versus averted gazes. This raises the possibility that gaze direction might differentially influence the performance of autistic individuals in a way that impacts the factor structure of the RMET.

1.5. Aims and Hypotheses

This study had three primary aims. The first aim was to evaluate the factor structure of the RMET and to test whether inconsistencies in previous factor analyses were due to differences in the way individuals with higher versus lower levels of autistic traits completed the task. I tested two hypotheses related to this aim:

H1a: The RMET is a multidimensional measure of ToM ability, and an EFA will reveal a multifactorial structure.

H1b: There is a different factor structure for the RMET amongst participants with higher levels of autistic traits compared to those with lower levels of autistic traits. As a consequence, running factor analyses separately on the data from the top and bottom

third of participants based on a standardised measure of autistic traits will result in better factor model fit for both groups compared to the fit for the overall sample.

The second aim was to evaluate Oakley et al.'s (2016) claim that autistic individuals score lower on the RMET due to difficulties with emotion recognition related to co-occurring alexithymia rather than a ToM deficit. I tested three hypotheses related to this aim:

H2a: Levels of autistic traits (as measured by the Autism Spectrum Quotient – brief version (AQ-28, Hoekstra et al., 2011)) and alexithymia traits (as measured by the TAS) will both negatively correlate with RMET scores.

H2b: After controlling for levels of alexithymia, autistic traits will not exhibit a statistically significant association with scores on the RMET.

H2c: In contrast, autistic traits, but not alexithymia traits, will exhibit a statistically significant positive association with ToM ability as assessed using a task that requires false belief understanding rather than relying solely on emotion recognition.

Previous quantitative and qualitative research has shown that some autistic individuals find eye contact and eye stimuli stressful (Hadjikhani et al., 2017; Kylliäinen & Hietanen, 2006; Trevisan et al., 2017). The third aim was to evaluate whether a self-report measure of the level of comfort participants feel when viewing the eye stimuli in the RMET predicted performance on the task. I tested three hypotheses related to this aim:

H3a: Reported comfort viewing the eye stimuli stressful will negatively correlate with performance on the RMET.

H3b: Reported comfort viewing the eye stimuli will negatively correlate with AQ-28 scores.

H3c: Extent to which AQ-28 score predicts RMET scores will be mediated by reported comfort viewing the eye stimuli.

As part of this study, participants completed the TAS and the AQ-28. Because the validity of measures can vary across different administration formats and different populations (Furr, 2011), the secondary aim of this study was to evaluate the factor structure of these two measures when

administered online to a large representative sample of the US population. I tested two hypotheses related to this aim:

H4: The TAS has three factors: difficulties identifying feelings, difficulties describing feelings, and externally oriented thinking. CFA will confirm the subscales of the TAS.

H5: The AQ-28 shows two higher-order factors 1) Social behaviour, which consists of four subscales (*social skills*, *routine*, *switching*, and *imagination*) and 2) Numbers and patterns. CFA will confirm the factor structure of the AQ-28.

2. Method

This study was pre-registered; however, there were some deviations from the pre-registration. Deviations are noted in the text and exploratory analyses are reported as such. The pre-registration, data, R scripts, and supplementary materials are available on the project's OSF page.[1]

2.1 Participants

Participants were recruited using Lucid Theorem (Coppock & McClellan, 2019), an online recruitment platform that uses quota sampling to provide a sample that matches the US national distribution in terms of age, gender, ethnicity, and geographic region. Lucid Theorem charges researchers US $1 per 15-minute survey completion and pays partner companies to supply research participants. Participants were compensated directly with cash, gift cards, or loyalty reward points by Lucid's partner companies according to the terms of their agreements with these partner companies. There were three pre-registered exclusion criteria. First, for analyses involving gender, only binary gender categories were analysed. No participants were excluded based on this criterion as Lucid Theorem provided binary gender classifications for all participants. Second, 863 participants who failed an attention check question in which they were asked to show that they have read the questions by moving a slider to "0", were excluded from all analyses. Third, 39 participants who did not finish the Imposing Memory Task (IMT), were excluded from the analyses involving this measure.

[1] Project's OSF page: https://osf.io/8jtn9/?view_only=447c1cbd822343559cfc7561ef444e0f.

At this stage, examination of the data revealed that 13 participants provided straight lined responses to the AQ-28, which means that they selected the same response across all items on this measure. Because approximately half of the items are reversed scored, it is extremely unlikely that these responses represent genuine attempts. A subset of participants also straight lined their responses to the TAS and IMT. In total, 41 participants provided straight line responses to one or more measures. Because these responses are unlikely to represent genuine attempts, these participants were excluded. While this was not a pre-registered exclusion criterion, it is in line with the exclusion criteria used by Olderbak et al. (2015). Where these additional exclusions result in changed patterns of results, this is noted in the text. In addition, results with these participants retained can be viewed on the project's OSF page.

The final sample with straight line responders removed included 1,181 (652 female) participants. Participants' ages ranged from 18-88 (*M* = 47.7, *SD* = 17.0). The sample was representative of the US population in levels of educational attainment (high school 23.5% [USA 28.3%], some university study 20.1% [17.7%], two year degree 8.4% [9.8%], four year degree 25.9%, [21.2%], postgraduate studies 13.2% [10.8%]) and race (White 76.6% [76.3%], Black 9.3% [13.4%], Asian 2.7% [5.9%], Other or prefer not to answer 13.0%).

Forty-two participants did not complete the IMT. As noted in the pre-registration, these participants were excluded from the analyses of the best predictors of RMET performance.

2.2. Measures

2.2.1. Reading the Mind in the Eyes Test (Revised Version)

The revised RMET (Baron-Cohen, Wheelwright, Hill et al., 2001) was designed to measure ToM ability in adults and comprises 36 items. Each item includes a black and white image of a person's eyes and four mental state terms. Participants are instructed to select "the word that best describes what the person in the picture is thinking or feeling."

The original test was in a paper format with one test item per page and the mental state terms printed around the four corners of the image (e.g. flustered, convinced, desired, joking). To

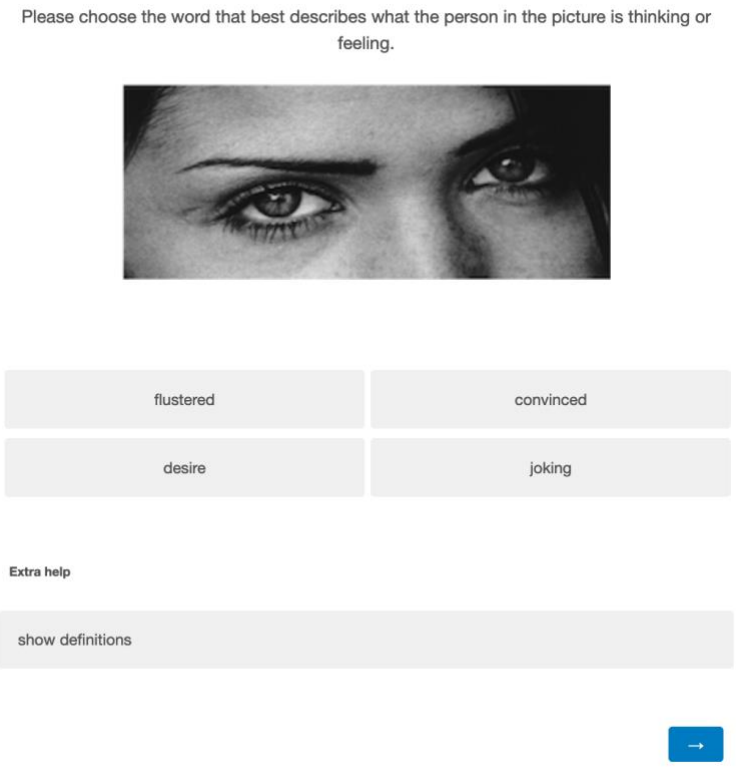
avoid any response bias, I presented the four mental state terms below the image with the word order randomised across participants (see Figure 2.).

The paper version of the test includes a glossary of terms to ensure that participants know the meaning of all the mental state terms. For each RMET question, I included an "extra help" section that participants could click to see the definitions of the mental state terms for that test item. Items are scored with 1 point for a correct response and 0 points for an incorrect response. The total score ranges from 0-36. Higher scores purportedly indicate higher levels of ToM ability.

Each test item was presented on a separate page, and the order of presentation of the items was randomised across participants.

Figure 2

Sample RMET Item Layout Used in this Study



Note. Image retrieved from the Autism Research Centre (2020) https://www.autismresearchcentre.com/tests/eyes-test-adult/

2.2.2. Autism Spectrum Quotient (Brief Version)

The AQ-28 (Hoekstra et al., 2011) is a 28-item reduced version of Baron-Cohen, Wheelwright, Skinner et al.'s (2001) 50-item Autism Spectrum Quotient questionnaire that is designed to measure autistic traits. Hoekstra et al. (2011) found that the AQ-28 has acceptable internal consistency (Cronbach's α = .78) and correlates highly with the 50-item scale (r = .93).

Each item consists of a statement and a four-point Likert scale on which participants rate how well each statement applies to them, from "strongly agree" to "strongly disagree." Hoekstra et al. (2011) identified a five-factor structure and a two-factor higher-order factor structure (see Table 3). For best model fit, one item was allowed to cross load on to both the social skills and routine factors.

Each item is scored from 1-4. Scores range from 28-112. Higher scores indicate more autistic traits. Items were presented in a randomised order across four screens, with seven items per screen.

Table 3

Factor Structure of the AQ-28

Higher Order Factor	Factor	Sample Item
Social behaviour		
	Social skills	I find social situations easy†
	Routine	New situations make me anxious
	Switching	I frequently get strongly absorbed in one thing
	Imagination	I find making up stories easy†
Numbers and patterns		
	Numbers and patterns	I am fascinated by dates

Note. † indicates items that are reverse scored

2.2.3 Toronto Alexithymia Scale

The TAS (Bagby et al., 1994) is a self-report measure of alexithymia that comprises 20 statements. Respondents rate how well each statement applies to them on a five-point Likert scale ranging from "strongly agree" to "strongly disagree." The TAS has three subscales, difficulty identifying feelings (*identify,* e.g. "I am often puzzled by sensation in my body"), difficulty describing feelings (*describe,* e.g. "I find it hard to describe how I feel about people"), and externally oriented thinking (*external,* e.g. "I prefer talking to people about their daily activities rather than their feelings").

Bagby et al. (1994) reported acceptable levels of test re-test reliability (.77) and internal consistency (with the exception of the *external* subscale: Cronbach's alpha for full test = .81, *identify* = .78, *describe* = .75, *external* = .66. Bagby et al. (2014) compared results from the TAS administered in a paper versus online version and reported similar levels of internal consistency for both the online (Cronbach's alpha for full scale = .80, *identify* = .82, *describe* = .71, *external* = .55) and paper versions (full scale = .75*, identify* = .75, *describe* = .69, *external* = .48).

Each item is scored from 1-5. Total scores range from 20-100, with higher scores indicating higher levels of alexithymia traits. Questions were presented in a randomised order across three screens with seven items on the first two screens and six items on the third screen.

2.2.4. Measure of Comfort Looking at Eye Stimuli

On the screen immediately following the final RMET test item, participants were asked to rate their level of comfort looking at the images of the eyes in the RMET. This was a single item measure of comfort designed for this study. Participants were presented with a slider with values from 1-10 and the following instruction: "The previous section involved looking at images of people's eyes. Move the slider below to indicate how comfortable you felt looking at the images of the eyes on a scale from 0 (very uncomfortable) to 10 (very comfortable)." Higher scores indicate higher levels of comfort viewing the eye stimuli.

2.2.5. Imposing Memory Task

The IMT was originally created by Kinderman et al. (1998) to explore the relationship between ToM deficits and causal attributions. They created their own ToM task because existing ToM tasks that assess causal mental-state reasoning confounded ToM ability with the ability to make causal attributions. The task has since been adapted by a number of researchers for use as a measure of higher-order ToM abilities in non-clinical populations (Launay et al., 2015; Stiller & Dunbar, 2007). Higher-order ToM relates to the ability to infer how people perceive the mental states of others across multiple iterations. For example, rather than determining what John thinks, a second-order ToM task would require determining what John thinks that Mary thinks, and a third-order ToM task would involve determining what John thinks that Mary thinks that Michael thinks.

Timing constraints for the overall survey completion time made it essential to select a ToM task that would provide a meaningful spread of possible scores without being overly time consuming to complete. Because it was anticipated that a large number of participants would complete the survey on a mobile device, it was also critical to minimise responses requiring text entry. Despite being a less widely used measure of ToM ability, similar to the more widely used Strange Stories task, the IMT is a story-based ToM task that also satisfied these constraints as outlined below.

For this study, I used a single story of approximately 200 words from Launay et al.'s (2015) version of the IMT (adapted from Stiller & Dunbar, 2007). While generally similar to the Strange Stories task, participants read the story and then answered true/false questions based on its content rather than being asked to generate an explanation that probes mental-state reasoning. Some of the true/false questions required mental state reasoning (e.g. "Carolyn thought that Hannah liked Emma's boyfriend Matt"), whereas others were memory control questions (e.g. "Carolyn told Hannah that Emma had been at training"). I used a subset of 16 of the 22 questions related to the story, comprising eight ToM questions and eight memory control questions. Questions scored 1 point for a correct response and 0 points for an incorrect response. Both memory and ToM scores were calculated with a range of 0-8 for each category.

I made two small amendments to the questions by replacing a pronoun with a character's name to reduce ambiguity and replacing the word "friend" with the word "colleague" in another question as it was not clear from the story that the two characters were friends. Materials for this task can be found in Appendix A.

2.3. Procedure

Participants completed a Qualtrics survey online in a location of their choice on an internet connected device (mobile phone 43%, laptop computer 26%, desktop computer 19%, tablet 11%). While the survey was intended to be completed in a single session, participants were able to leave and return to the survey. Duration was recorded based on the time the survey was initially opened until the time the final question was submitted. The median completion time was 18 minutes. Data collection occurred from August 25th through September 10th, 2020.

After the consent page, the RMET, TAS, and AQ-28 were presented with their order randomised across participants. The question about comfort viewing eye stimuli was always presented immediately after the RMET. Because the IMT required a greater level of reading comprehension, it was presented after the other measures to avoid participant fatigue. The survey finished with demographic questions. The demographic questions included a measure of belief in God that is part of a separate analysis, and the results will be reported elsewhere. Lucid Theorem also supplied demographic statistics, which were used for reporting and analysing of demographic information. Ethics approval for this study was granted by Macquarie University Human Research Ethics Committee (Appendix B, reference number: 52020625515320).

2.4. Analytic Approach

In my analyses, model fit was assessed using seven metrics: Chi-square, root mean square of residuals (RMSR), standardised root mean square of residuals (SRMR), root mean square error of approximation (RMSEA), comparative fit index (CFI), Tucker-Lewis index (TLI), and Bayesian information criterion (BIC). A non-significant chi-square indicates good model fit, however, as sample sizes increase, chi-square becomes significant independent of model fit (Bergh, 2015). I have

reported chi-square for all models, however, because my sample size is very large, chi-square was always significant. CFI and TLI are relative fit measures, which means that they compare model fit to a null model. Higher values indicate better model fit. In contrast, RMSR, SRMR, and RMSEA are absolute fit measures, and model fit is evaluated without comparison to a null model. Lower values indicate better model fit. Lower values also indicate better model fit for BIC, and the value can be used to select between competing models.

The most frequently reported measure of internal consistency is Cronbach's alpha, however, recently coefficient omega has been recommended as a more appropriate measure of internal consistency due to psychometric limitations of alpha as an indicator of reliability (Dunn et al., 2014; Goodboy & Martin 2020). There are a variety of ways in which coefficient omega can be calculated, and the most appropriate calculation method varies according the characteristics of the test being evaluated. In line with the recommendations of Flora (2020), omega hierarchical was calculated for the RMET using the *omega* function in the *psych* package (v.2.0.9, Revelle, 2020) in R (version 4.0.1, 2020-06-06, The R Core Team, 2020), whereas omega hierarchical[2] for both the TAS and AQ-28 were calculated with the *reliability* function in the *semTools* package (Jorgensen et al., 2020) in R.

2.4.1. RMET Factor Structure (H1a and H1b)

EFA was used to evaluate the factor structure of the RMET. An item was considered to load onto a factor if the rotated factor weighting was ≥ 0.3. Model fit was evaluated against the following criteria: RMSR < .05, RMSEA < .08 acceptable fit, < .05 good fit, TLI ≥ .95, CFI ≥ .90, chi-square, $p > .05$, and BIC (lower values indicating better model fit); however, these values are guidelines rather than established strict cutoff criteria (Marsh et al., 2004). In addition to evaluating models according to these metrics, decisions were made based on the conceptual applicability of the models.

[2] Omega hierarchical is called omega3 in the R output.

2.4.2. Best RMET Predictors (H2a-H2c)

To determine the best predictors of RMET performance, I used structural equation modelling (SEM) and an exploratory hierarchical regression analysis to evaluate the relations between RMET performance and TAS scores, AQ-28 scores, and IMT scores, while controlling for gender (gender has been shown to be associated with performance on the RMET in individuals without an autism diagnosis, Baron-Cohen et al., 2015). Model fit for all SEM analyses was evaluated by model chi-square, $p > .05$, CFI ≥ .90, TLI ≥ .95, RMSEA < .06, and SRMR < .08.

2.4.3. Relationship Between Comfort Viewing Eye Stimuli, AQ-28, and RMET (H3a-H3c)

To test whether comfort viewing the eye stimuli is related to RMET performance, I conducted linear modelling to evaluate whether performance on the RMET correlated positively with reported comfort looking at the stimuli (H3a) and AQ-28 scores correlated negatively with reported comfort looking at the eyes (H3b). I used mediation analysis to evaluate whether there was an indirect effect between AQ-28 scores and RMET scores that was mediated by comfort viewing the eye stimuli (H3c).

2.4.4. CFA on the TAS and AQ-28 Subscales (H4 and H5)

CFA was used to evaluate the proposed factor structures of the TAS (H4) and AQ-28 (H5) in an online US representative sample. Model fit was evaluated with the SEM fit indices indicated above.

3. Results

3.1. Descriptive Statistics

Table 4 contains descriptive statistics for the outcome variables, and Table 5 contains the correlation matrix for all variables. Appendix C contains histograms of RMET, AQ-28, and TAS scores. RMET scores were normally distributed (skew = -0.73, kurtosis = 0.35). Scores ranged from 5 (14% correct) to 34 (94% correct). Average scores (M = 23.49, 65% correct, SD = 5.51) were lower than scores in the general population reported by Baron-Cohen, Wheelwright, Hill et al. (2001, M = 26.2 SD = 3.6) and Olderbak et al. (2015, M = 27.2, SD = 3.82). There were eight test items that failed

Baron-Cohen and colleagues' initial criteria for validating the test items: greater than 25% of participants selected the same foil for items 6, 10, 17, 23, 25, 28, 34, and 35 (see Appendix D). Less than half of participants selected the correct response for two of these items (23, 25). Cronbach's alpha was .75, however, as discussed above, this value is likely inflated due to the length of the measure. Internal consistency was poor according to omega hierarchical (.59).

Table 4

Descriptive Statistics for Outcome Variables

Measure	Min	Max	Range	Mean	SD
RMET	5	34	0-36	23.5	5.5
AQ-28 total	39	102	28-112	66.0	9.5
AQ-28 *imagination*	8	30	8-32	17.2	4.0
AQ-28 *social skills*	8	32	8-32	19.4	5.2
AQ-28 *switching*	4	16	4-16	9.4	2.2
AQ-28 *routine*	4	16	4-16	10.3	2.4
AQ-28 *numbers*	5	20	5-20	12.5	3.3
TAS total	22	83	20-100	49.4	12.3
TAS *describe*	5	25	5-25	13.1	4.7
TAS *identify*	7	35	7-35	16.0	6.5
TAS *external*	8	38	8-40	20.3	4.5
IMT memory	0	8	0-8	6.0	1.6
IMT ToM	0	8	0-8	5.8	1.6
Comfort	0	10	0-10	7.3	2.5

Table 5

Correlation Matrix for all Variables

Variable	Age	RMET	TAS total	TAS Describe	TAS Identify	TAS External	AQ-28 Total	AQ-28 Social Behaviour	AQ-28 Social Skills	AQ-28 Imagination	AQ-28 Switching	AQ-28 Numbers	AQ-28 Routine	Comfort	IMT Memory
Age															
RMET	0.18***														
TAS Total	-0.23***	-0.20***													
TAS Describe	-0.21***	-0.08**	0.85***												
TAS Identify	-0.29***	-0.18***	0.86***	0.65***											
TAS External	0.01	-0.20***	0.62***	0.35***	0.24***										
AQ-28 Total	-0.09	-0.01	0.45***	0.43***	0.38***	0.24***									
AQ-28 Social Behaviour	-0.04	0.00	0.44***	0.42***	0.33***	0.29***	0.94***								
AQ-28 Social Skills	-0.08	0.11*	0.32***	0.35***	0.26***	0.14***	0.80***	0.83***							
AQ-28 Imagination	0.03	-0.16***	0.37***	0.30***	0.22***	0.38***	0.62***	0.69***	0.28***						
AQ-28 Switching	-0.05	0.02	0.32***	0.29***	0.29***	0.15***	0.63***	0.67***	0.42***	0.38***					
AQ-28 Numbers	-0.12**	-0.05	0.03	0.03	0.14***	-0.17***	0.13**	-0.22***	-0.14***	-0.22***	-0.17***				
AQ-28 Routine	-0.06	0.05	0.31***	0.28***	0.27***	0.15***	0.69***	0.70***	0.59***	0.26***	0.41***	-0.06			
Comfort	0.12**	0.19***	-0.21***	-0.21***	-0.14***	-0.17***	-0.21***	-0.24***	-0.18***	-0.20***	-0.19***	0.11*	-0.11*		
IMT Memory	0.13***	0.37***	-0.17***	-0.10	-0.16***	-0.15***	-0.02	0 .00	0.07	-0.12**	0.00	-0.05	0.03	0.08	
IMT ToM	0.11*	0.34***	-0.16***	-0.09**	-0.14***	-0.13***	-0.04	-0.02	0.05	-0.11*	0.00	-0.07	0.01	0.10	0.51***

Note. * p < 0.05, ** p < .001, *** p < .001

The mean inter-item tetrachoric correlation for the RMET was 0.13 (range from –0.12 to 0.36, see Appendix E for the full correlation matrix) which, consistent with previously reported values (.10, Black, 2019; .08, Oakley et al., 2016), falls below the range recommended by Clark and Watson (1995) of 0.15-0.50 and indicates low levels of agreement between items. While Clark and Watson note that broader, higher-order constructs (such as ToM) may have mean inter-item correlations near the bottom of this range, they also recommend that the majority of individual inter-item correlations fall within the 0.15-0.50 range. In my data, only half of the inter-item correlations fall within this range. Also in line with Olderbak et al. (2015), the correlations between items with same target was low (fantasizing r = .21, cautious r = .10, preoccupied r = .32, interested r = .15).

The TAS was normally distributed (skew = 0.13, kurtosis = -0.65). The mean score on the TAS was 49.4 (*SD* = 12.3). Internal consistency was within an acceptable range for the full scale (ω_h = 0.87) and the TAS *describe* and TAS *identify* subscales (ω_h = .78, ω_h = .85, respectively). Consistent with Bagby et al. (1994, 2014), the internal consistency of the TAS *external* subscale was low (ω_h = .47).

AQ-28 scores were normally distributed (skew = 0.24, kurtosis = 0.27). Scores ranged from 39-102 (*M* = 66.0, *SD* = 9.5). Internal consistency was within the acceptable range for the full scale (ω_h = .80) and the *social skills* (ω_h = .71) and *numbers and patterns* subscales (ω_h = .72). However, the reliability of the other three subscales were below the recommended range (*routine*, ω_h = .46, *switching,* ω_h = .57, *imagination*, ω_h = .66).

3.2. EFA of Overall RMET Factor Structure (H1a)

For the full sample, I ran parallel analysis to determine the number of factors to retain, using the *psych* package (version 2.0.9, Revelle, 2020) in R. Because the items are dichotomous, I used a tetrachoric correlation matrix. I used weighted least squares factoring method with an oblique rotation (geominQ) and 50 iterations, as this is an appropriate method to use with dichotomous data (Susana et al., 2017). Parallel analysis suggested 12 factors, and model fit measures for this solution

approached good fit levels (CFI: .889, TLI .732, RMSEA = .051 [.045-.055], RMSR = .02, BIC = -782, chi-square = 1086, p < 0.001). However, in this model, eleven items failed to load on to any factor, seven factors consisted of only a single item, and the other factors lacked any obvious conceptual explanatory power.

As an exploratory analysis, I applied the comparative fit method used by Olderbak et al. (2015). Starting with a single-factor model, I increased the number of factors until I had satisfactory model fit as assessed by RMSEA, CFI, and TLI. Model fit improved with every additional factor retained, but relative fit statistics did not approach satisfactory levels until a model with 14 factors (see Appendix F). Similar to the 12-factor model, many items failed to load on to any factor in the 14-factor model, 12 factors had only one or two items, and factors lacked conceptual clarity.

While Parallel analysis is considered to be one of the most reliable methods to determine the number of factors to maintain in EFA (Hayten et al., 2004), because there was a poor conceptual fit for the model retaining the number of factors indicated by parallel analysis, I conducted an exploratory analysis using Cattell's scree plot[3] (see Figure 3). This approach indicated a three-factor solution. This model had acceptable fit when evaluated by RMSEA, .059 (.057-.061) and good model fit according to RMSR (.05) and chi-square (2693, p <0.001.) However, model fit was poor according to global fit indices (CFI = .706, TLI .647, BIC = -1021). Similar to the models with more factors retained, there were nine items that did not load onto any of the three factors, three items had cross loading on two factors, and the maximum factor loading was only 0.55 (see Table 5). The cumulative variance explained was .20.

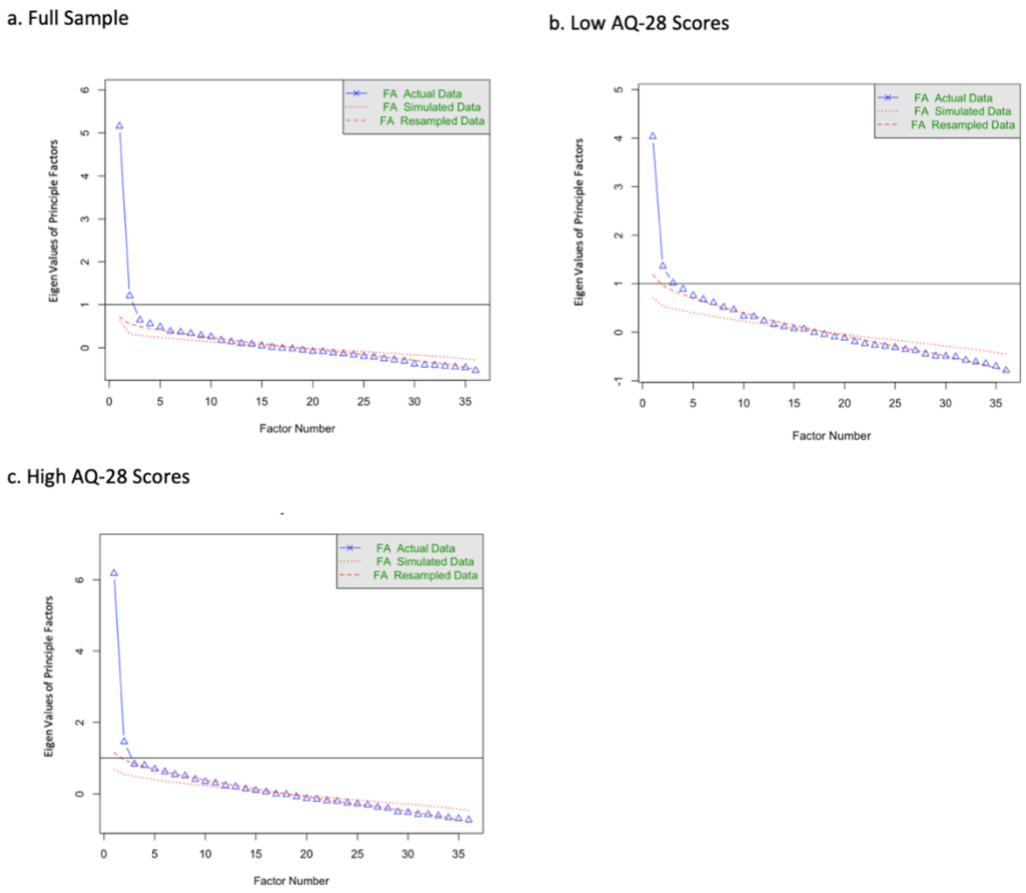
Despite the poor model fit and high number of items failing to load onto any factor, the three-factor solution did have conceptual explanatory power, with one factor relating to internally

[3] Another method of determining how many factors to retain is to retain only factors with eigenvalues ≥ 1. This method indicated a two-factor solution. The three-factor model was preferred because it had better model fit indices and better conceptual explanatory power, however both the two and three-factor models suffered from the same limitations of overall poor model fit, low factor loadings, and a high number of factors failing to load on to any factor, which indicated that ultimately, both models should be rejected.

oriented attention and thinking (e.g. pensive, preoccupied), one factor relating to negative emotions (e.g. hostile, despondent), and one factor relating to flirtation (e.g. flirtatious, fantasizing). The flirtatious factor overlaps with one of the five factors identified by Olderbak et al. (2015), with five items in common (desire (3), flirtatious(30), fantasizing (21,6), interested (25)). However, a number of items failed to load as would be expected. The target "reflective" (29) did not load on to the thoughtful factor. The targets "upset" (2), "worried" (5), and "accusing" (14) did not load on to the negative factor. Only one of the two RMET items with the target "interested" loaded on to the flirtatious factor.

Figure 3

Scree Plots for Parallel Analysis of RMET Data



Note. FA = Factor Analysis

Table 6

Factor Loadings for Three Factor EFA on the Full Sample[4]

RMET item	Target	Factor 1 Thoughtful	Factor 2 Negative	Factor 3 Flirtatious
32	serious	0.514		
22	preoccupied	0.488		
24	pensive	0.486		
9	preoccupied	0.460		
17	doubtful	0.412		-0.408
15	contemplative	0.411		
16	thoughtful	0.396		
20	friendly	0.368		
27	cautious	0.343		
33	concerned	0.319		
28	interested	0.304		
4	insisting		0.490	
8	despondent		0.442	
26	hostile		0.440	
7	uneasy		0.411	
34	distrustful		0.407	
12	sceptical		0.361	
11	regretful		0.352	
35	nervous		0.310	
36	suspicious		0.306	
23	defiant		0.303	
30	flirtatious		0.498	0.550
21	fantasizing	0.327		0.462
3	desire			0.408
6	fantasizing			0.381
25	interested			0.351
1	playful			0.334

[4] Running this analysis on the full dataset (i.e., including straight lining participants) led to a slightly different factor structure: item 23 no longer loaded on the negative factor and item 21 no longer loaded on the thinking factor.

Because the RMET is proposed to be a single factor test, I conducted an exploratory analysis using CFA to evaluate a single factor model, despite EFA results suggesting poor model fit for a single factor (CFI = .580, TLI = .554, RMSEA = .066 [.064-.069], RSMR = .06, BIC = -506, χ^2 = 3697, p < .001, Highest factor loading = 0.62, Items failing to load on to the factor = 7). Notably, five of the seven items that failed to load on to the single factor in the EFA also failed Baron-Cohen's original criteria for inclusion in the RMET because more than 25% of participants selected the same incorrect foil. For the CFA, robust fit indices showed good fit according to absolute fit indices (RMSEA = .025 [.022-.027], SRMR = .036, χ^2 [594, N = 1181] = 1029, p < 0.001), but poor fit according to relative fit indices (CFI = .865, TLI = .856).

3.3. EFA of RMET Factor Structure in Participants Low in Autistic Traits (H1b)

To evaluate the possibility that individuals with lower levels of autistic traits use different strategies to complete the RMET, I conducted EFA on the top and bottom third of participants based on AQ-28 scores. Scores for participants in the bottom third ranged from 39-62 (*M* = 56.2, *SD* = 4.8). This group consisted of 422 (233 female) participants with a mean age 49.2 (*SD* = 16.1). This group was demographically similar to the full sample (White 78%, Black 11%, Asian 3%, other or prefer not to answer 6%). The mean RMET score was 24.2 (*SD* = 4.7). Internal consistency was below the acceptable range (Cronbach's α = .68, ω_h = .40), and some items negatively correlated with the scale.

Parallel analysis suggested 13 factors, however, the 13-factor model did not converge. In fact, no factor solution from 1-13 resulted in good model fit. Considering this, I used Cattell's scree plot, which indicated a three-factor solution (see Figure 3). All fit measures indicated poor model fit (CFI = .430, TLI = .312, RMSEA = .094 [.090-.098], RMSR = .07, BIC = -704, χ^2 = 2469, p < .001.). Thirteen items did not load onto any factor and three items had cross-loadings (see Table 6). The three factors overlapped considerably with the results from the full sample, and conceptually matched the division into thoughtful, negative, and flirtatious factors. The maximum factor loading of .615 was higher than for the full sample. The cumulative variance was comparable to the full sample: .19.

A number of items failed to load as expected. For example, distrustful (34) and sceptical (12), which loaded on to the negative factor in the full sample loaded on to the thoughtful factor in the low AQ-28 scores group, and fantasizing (6) no longer loaded on to the flirtatious factor.

There were insufficient data points to run a CFA for single factor model within this subset of the data for the full 36-item test.

Table 7

Factor Loadings for Participants with Low AQ-28 Scores[5]

RMET item	Target	Factor 1 Thoughtful	Factor 2 Negative	Factor 3 Flirtatious
16	thoughtful	0.603		
24	pensive	0.526		
29	reflective	0.414		
9	preoccupied	0.404		
22	preoccupied	0.401	0.343	
5	worried	0.358		
14	accusing	0.356		
28	interested	0.350		
13	anticipating	0.346		
34	distrustful	0.325		
12	sceptical	0.310		
4	insisting		0.579	
26	hostile		0.390	
36	suspicious		0.379	
8	despondent		0.363	
27	cautious		0.352	
11	regretful		0.343	
35	nervous		0.324	
3	desire			0.615
21	fantasizing			0.585
30	flirtatious	0.310		0.427
25	interested			0.417
31	confident			0.384

[5] Running this analysis on the full dataset (i.e., including straight lining participants) led to a slightly different factor structure: item 18 loaded on to the thoughtful factor loaded on to both the thinking and negative factors.

3.4. EFA of RMET Factor Structure in Participants High in Autistic Traits (H1b)

Scores for the group of participants with AQ-28 scores in the top third ranged from 70-102 (*M* = 76.1, *SD* = 5.9). Hoekstra et al. (2011) reported that a cut-off scores of ≥ 70 had a specificity of .91 and sensitivity of .94 in distinguishing autistic individuals from controls. It is unlikely that 30% of the participants in my US representative sample would qualify for an autism diagnosis. Although my sample is demographically representative, it is possible that participants who complete surveys online for money are more inclined to the traits measured by the AQ-28. This group consisted of 409 (234 female) participants with a mean age of 45.7 (*SD* = 17.0). Demographics were similar to the full sample (White 74%, Black 10%, Asian 3%, other or prefer not to answer 10%). Mean RMET score for this group was 23.5 (*SD* = 5.7). Internal consistency of the RMET was acceptable according to Cronbach's alpha (.78) but low according to omega hierarchical (.57).

Parallel analysis on the data of participants with high AQ-28 scores suggested 14 factors, however the 14-factor model did not converge. Instead, I used Cattell's scree plot method, which indicated a three-factor model (see Figure 3). Model fit was poor across all fit statistics (CFI = .486, TLI = .379, RMSEA = .107 [.104-.111], RMSR = .07, BIC = -154, χ^2 = 3003, $p < .001$). Eight items did not load on to any factor, and four items had cross loadings on two factors (See Table 7). This group had the highest maximum factor loading (0.807). The factor structure for the high AQ-28 group was notably different from the full sample and the low AQ-28 group. Twenty-one items loaded on to Factor 1. Factor 2 only had three items, and two of these items had cross loadings. Factor 3 only had four items, one of which had a cross loading on Factor 1. All three factors contained stimuli with both direct and averted gaze. Unlike the full sample and low AQ-28 group, the removal of the straight lining participants impacted the factor structure for this group considerably (see Appendix G for a comparison of the two factor model results). The cumulative variance explained was 0.25.

As with the low AQ-28 group, there were insufficient data points to run a CFA for single factor model.

Table 8

Factor Loadings for Participants with High AQ-28 Scores

RMET item	Target	Factor 1	Factor 2 Negative	Factor 3 Thoughtful
30	flirtatious	0.807		
21	fantasizing	0.766	-0.405	
3	desire	0.509		
9	preoccupied	0.502		
32	serious	0.498		
25	interested	0.452		
13	anticipating	0.442		
26	hostile	0.441		
8	despondent	0.440		
29	reflective	0.437		
36	suspicious	0.428		
6	fantasizing	0.405		
1	playful	0.405		
31	confident	0.405		
16	thoughtful	0.400		
20	friendly	0.382		
15	contemplative	0.369		
18	decisive	0.367		
12	sceptical	0.362		
2	upset	0.358		
5	worried	0.329		
7	uneasy		0.430	0.329
34	distrustful	0.369	0.413	
4	insisting		0.326	
22	preoccupied			0.583
28	interested			0.528
24	pensive	0.357		0.374
17	doubtful			0.346

3.5. Best RMET Predictor Variables (H2)

3.5.1. Structural Equation Modelling of RMET Predictors

In line with the pre-registration, 39 participants who did not complete the IMT were excluded from this analysis, leaving 1142 (624 female) participants.

Because the primary question of interest was whether the AQ-28 and TAS scores are significantly associated with RMET scores after controlling for other variables (H2a, H2b), I tested a SEM model with paths to the RMET from gender, the three TAS subscale scores, the five AQ-28 first-order subscale scores, IMT memory scores, and IMT ToM scores. The resulting model had 0 degrees of freedom, indicating a saturated model. This means that model fit could not be evaluated, and the SEM resulted in a multiple linear regression of RMET scores on the variables (see Table 8).

Table 9

SEM RMET Regression Results

Variables	*b*	SE(*b*)	*β*
Gender	0.690	0.290	0.064*
AQ-28 *imagination*	-0.159	0.042	-0.119***
AQ-28 *social skills*	0.110	0.036	0.106**
AQ-28 *switching*	0.121	0.075	0.051
AQ-28 *routine*	0.041	0.075	0.018
AQ-28 *numbers*	-0.026	0.046	-0.016
TAS *describe*	0.102	0.042	0.090*
TAS *identify*	-0.139	0.029	-0.170***
TAS *external*	-0.140	0.036	-0.117***
IMT memory	0.723	0.101	0.220***
IMT ToM	0.576	0.102	0.172***

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

This analysis indicated that all three TAS subscales and the AQ-28 *imagination* and *social skills* subscales correlated with RMET scores. I conducted an additional exploratory analysis in order to evaluate the independent contributions of the variables that had a significant relationship with RMET scores according to the SEM, using hierarchical regression analysis with ordinary least squares regressions.

3.5.2. Hierarchical Regression Analysis of RMET Predictors

Variables were entered in five steps: (1) gender (2) AQ-28 subscales of *social skills* and *imagination* (3) TAS subscales (4) IMT memory and ToM scores (5) comfort. The first three steps match the hierarchical regression analysis conducted by Oakley et al. (2016), except that in the current study subscales of the AQ-28 and TAS were used instead of the full scales. Because the SEM analysis also revealed significant relationships between the RMET scores and comfort and IMT, I conducted two additional steps to evaluate the contribution of these variables. ANOVA analysis was used to compare the significance of the successive models (see Table 9).

Each successive model resulted in a significant increase in R^2, indicating that all factors explain a portion of the observed variance in RMET scores. The correlations between variables and RMET scores were consistent with the results of the SEM in both direction and significance. The first model with only gender had an R^2 value of .015, indicating that gender accounted for approximately 1.5 % of the variation in RMET scores, with females performing better on average. With the addition of the AQ-28 *imagination* and *social skills* subscales in step 2, the variance accounted for increased to approximately 6%, and with the addition of the TAS subscales, it increased to approximately 11%. This indicates that the TAS and AQ-28 account for a similar proportion of the variance in RMET scores. R^2 doubled to .22 with the addition of the IMT scores. The addition of comfort resulted in a significant increase in R^2 of 0.02. While all of these variables were significantly correlated with RMET scores, the adjusted R^2 for the final model was only 0.24, indicating that this model accounts for less than a quarter of the variability in RMET scores.

Table 10

Hierarchical Regression Analysis Predicting Scores on the RMET

	Step 1			Step 2			Step 3			Step 4			Step 5		
Variable	*b*	*SE(b)*	*β*	*b*	*SE(b)*	*β*	*b*	*SE(b)*	*β*	*b*	*SE(b)*	*β*	*b*	*SE(b)*	*β*
Gender	1.331	.316	.124***	1.016	.312	.095**	.976	.304	.091**	.731	.286	.068*	.815	.282	.076**
AQ *imagine*				-.273	.040	-.204***	-.176	.042	-.131***	-.134	.040	-.100***	-.113	.039	-.084**
AQ *social*				.165	.031	.159***	.189	.032	.182***	.138	.030	.133***	.154	.030	.149***
TAS *external*							-.185	.037	-.155***	-.138	.035	-.116***	-.127	.035	-.106***
TAS *describe*							.112	.045	.099*	.102	.042	.090*	.124	.042	.108**
TAS *identify*							-.184	.030	-.225***	-.133	.029	-.163***	-.137	.028	-.167***
IMT memory										.728	.101	.221***	.716	.100	.217***
IMT ToM										.582	.103	.173***	.548	.102	.163***
Comfort													.316	.058	.148***
R^2		.015			.061			.113			.221			.241	
F for change in R^2					35.78***			27.28***			82.06***			30.13***	

Note. * p < 0.05, ** p < .001, *** p < .001

To evaluate whether the same variables underscored performance on an alternate ToM task (H2c), I also conducted exploratory hierarchical regression analysis for predictors of the IMT ToM task using the variables that significantly correlated with IMT ToM scores (see Table 10). The first model with only gender had an R^2 of .011, indicating that gender accounted 1.1% of the variability in IMT ToM scores, with females performing better than males on average. The addition of the IMT memory task in step 2 resulted in a large increase in R^2 (.27). The AQ-28 *imagination* subscale was significant in step 3; however, this model did not result in a significant increase in R^2, and this subscale was no longer significant after the addition of the TAS subscales. The TAS *identify* subscale was significant in step 4, however this model did not result in a significant increase in R^2, and this subscale was not significant after the addition of the RMET in step 5. The addition of RMET scores resulted in a significant increase in R^2, indicating that RMET scores accounted for approximately 2% of the variation in IMT ToM scores.

3.6. Comfort Viewing Eye Stimuli and RMET Scores (H3)

I used ordinary least squares regression and mediation analysis to test the hypothesis that comfort viewing eye stimuli is positively correlated with RMET scores (H3a), negatively correlated with AQ-28 scores (H3b) and mediates the relationship between AQ-28 scores and RMET scores (H3c). I regressed RMET scores on total AQ-28 scores and found that AQ-28 scores did not significantly predict RMET performance, $F(1,1179) = 0.25$, $p = .20$, $R^2 = 0.00$. I next regressed RMET scores on comfort scores while controlling for AQ-28 scores. The level of comfort viewing the eyes did predict RMET scores, $F(1, 1179) = 21.83$, $p < .001$, $R^2 = .04$. Finally, I regressed comfort on AQ-28 scores and found that AQ-28 scores were significantly correlated with comfort, $F(1,1179) = 51.82$, $p < .001$, $R^2 = 0.04$.

Table 11

Hierarchical Regression Analysis Predicting Scores on the IMT ToM

	Step 1			Step 2			Step 3			Step 4			Step 5		
Variable	*b*	*SE(b)*	*β*	*b*	*SE(b)*	*β*	*b*	*SE(b)*	*β*	*b*	*SE(b)*	*β*	*b*	*SE(b)*	*β*
gender	.355	.094	.113***	.261	.081	.082**	.255	.081	.080**	.256	.081	.080**	.205	.081	.064*
IMT memory				.494	.025	.504***	.490	.025	.499***	.478	.025	.487***	.427	.027	.436***
AQ imagine							-.017	.010	-.0430*	-.008	.011	-.021	-.004	.011	-.009
TAS external										-.015	.010	-.041	-.007	.010	-.020
TAS describe										.008	.012	.025	.001	.012	.003
TAS identify										-.017	.008	-.069*	-.010	.008	-.042
RMET													.048	.008	.160***
R^2		.011			.264			.267			.269			.289	
F for change in R^2					406.50***			2.96			2.57			33.34***	

Note. * p < 0.05, ** p < .001, *** p < .001

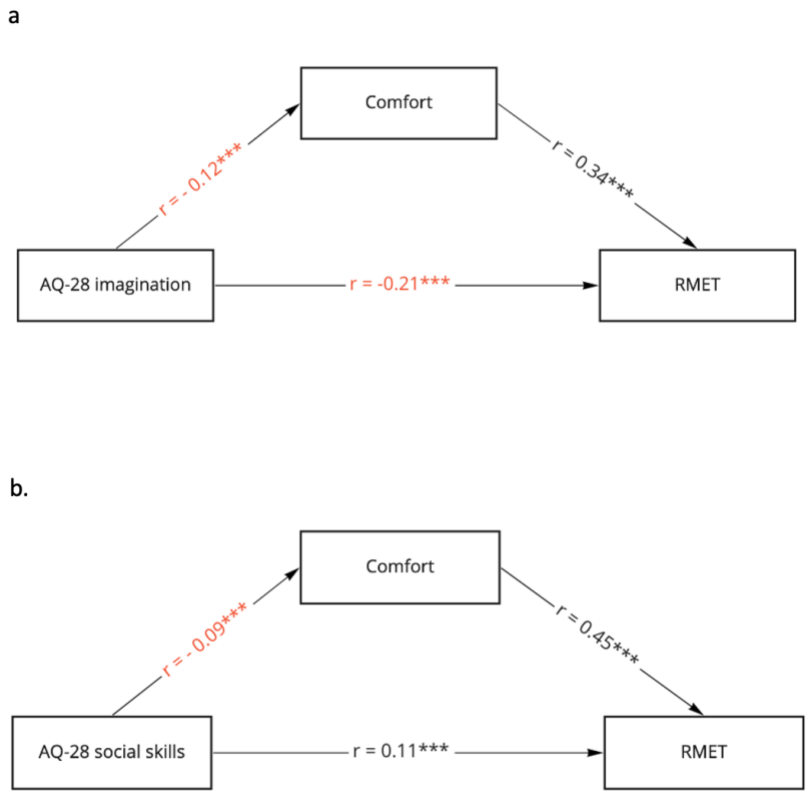
While AQ-28 total scores did not correlate with RMET scores, I conducted mediation analysis to see whether there was an effect that was entirely mediated via comfort. I ran a mediation analyses in R using the *mediation* package (v.4.5.0, Tingley et al., 2014). The average causal mediation effect was significant (-.020, $p < .001$), but the upper 95% CI was very close to zero (-.010). Consistent with the linear modelling results, there was no significant direct effect between AQ-28 and RMET scores (.014, $p = .370$). The total effect, which is the sum of the direct effect and mediated effect was not significant (-.001, $p = .570$), and the proportion of the effect that was mediated was not significant (2.702, $p = 0.570$), indicating that there is not a significant effect of AQ-28 scores on RMET scores that is mediated by comfort.

Two subscales of the AQ-28 were associated with the RMET (see Table 4). The *social skills* subscale had a weak positive correlation ($r = .11$, $p < .001$) with RMET scores, and the *imagination* subscale had a weak negative correlation with RMET scores ($r = -.16$, $p <.001$). I ran mediation analyses to see whether comfort viewing the eyes mediated the relationship between RMET scores and either the *social skills* or *imagination* subscales. First, I used linear regression to determine the relationship between the subscales and the RMET, the subscales and comfort, and comfort and RMET scores controlling for AQ-28 subscale scores. These relationships are presented in Figure 4.

For the *imagination* subscale the average causal mediation effect was significant and negative (see Table 11), which indicates that there is a negative correlation between the *imagination* subscale and RMET scores that is mediated by comfort. Consistent with the results from the linear modelling, the average direct effect of *imagination* subscale on RMET scores was also negative, indicating that higher levels of autistic traits related to imagination correlated with lower RMET scores. The total effect, which is the sum of the mediation and direct effects was also significant. The total effect was of a greater magnitude than the direct effect for *imagination* because both the direct and indirect effects were negative.

Figure 4

Diagrams of the Relationship between Comfort, RMET, and AQ-28 Subscale Scores



Note. (a) Comfort mediating the relationship between AQ-28 *imagination* subscale scores and RMET scores. (b) Comfort mediating the relationship between AQ-28 *social skills* subscale scores and RMET scores. Negative correlations are highlighted in red. *** = p < .001

Table 12

Mediation Analysis of Comfort, RMET, and AQ-28 Imagination Subscale Scores

	Estimate	95% CI Lower	95%CI Upper	p-value
ACME	-.043	-.064	-.020	<.001
ADE	-.170	-.247	-.100	<.001
Total Effect	-.213	-.290	-.140	<.001
Proportion mediated	.201	.111	.340	<.001

Note. ADE = direct effect between *imagination* and RMET scores, ACME = average causal mediation effect (via comfort), Total Effect is the sum of ADE and ACME.

The mediation was also significant for the *social skills* subscale (see Table 12). The average causal mediation effect was significant and negative, indicating that there was a significant mediating effect. The direct effect of *social skills* on RMET scores was positive, indicating that more difficulty with social skills is correlated with *higher* RMET scores. The total effect was significant, but unlike the *imagination* subscale, the magnitude of the total effect of *social skills* on RMET scores was lower than the direct effect because the direct effect was positive, whereas the indirect effect was negative.

Table 13

Mediation Analysis of Comfort, RMET, and AQ-28 Social Skills Subscale Scores

	Estimate	95% CI Lower	95%CI Upper	p-value
ACME	-.039	-.059	-.020	<.001
ADE	.152	.095	.200	<.001
Total Effect	.113	.060	.160	<.001
Proportion mediated	-.347	-.791	-.180	<.001

Note. ADE = direct effect between *imagination* and RMET scores, ACME = average causal mediation effect (via comfort), Total Effect is the sum of ADE and ACME.

3.7. CFA for TAS Subscales (H4)

Using CFA, I evaluated the three subscales of the TAS in a US representative sample completing the TAS online. Analyses were run using the *cfa* function in *lavaan* (v. 0.6-7, Rosseel, 2012). I found acceptable model fit according to absolute fit indices with the exception of chi-square (RMSEA = .058 [.055 – .062], SRMR = .059, χ^2 [167, N = 1181] = 841, $p < 0.001$). Bagby et al. (1994) also found had a significant chi-square, but they note that a positive chi-square can result from large sample sizes. Relative fit indices were just below the cutoff for acceptable model fit (CFI = .894, TLI = .879).

Twenty three participants provided straight lined responses to the TAS. Notably, even with our large sample size, the results of the CFA with the data from these straight line responders retained resulted in worse model fit (CFI = .824, TLI = .800, RMSEA = .071 [.068-.075], SRMR = .074, χ^2 [167, N = 1222] = 1203, p < 0.001) demonstrating the importance of data quality in online research.

3.8. CFA for AQ-28 Subscales (H5)

Using CFA, I evaluated the factor model of the AQ-28 proposed by Hoekstra et al., (2011), which includes five first-order factors and one second-order factor. I allowed item 26 ("new situations make me anxious") to load on to both the *routine* and *social skills* factors in line with Hoekstra et al. (2011). First, I evaluated a single order model excluding the second-order factor, social behaviour (see Table 1). Absolute fit indices indicated acceptable model fit (RMSEA = .054 [.051-.057], SRMR = .061), however relative fit indices indicated poor model fit (CFI = .821, TLI = .800, χ^2 [339, N = 1181] = 1490, p < .001). Model fit decreased across all fit indices with the addition of the second-order factor (CFI = .798 , TLI = .778, RMSEA = .057[0.054-0.060], RMSR = .066, χ^2 [345, N = 1181] = 1667, p < .001).

Thirteen participants provided straight lined responses to the AQ-28. As with the analyses of the TAS, model fit improved with the removal of participants with straight lined responses. (first-order model: CFI = .728, TLI = .697, SRMR = 0.073, RMSEA = 0.062 ([.059-.068], χ^2 [339, N = 1222] = 1926, p < 0.001, second order model: CFI = .698, TLI = .669, RMSEA = .065 [.062-.067], SRMR = .061, χ^2 [345, N = 1222] = 2109, p < .001).

4. Discussion

4.1. Factor Structure of the RMET (H1a and H1b)

One primary aim of this study was to evaluate the factor structure of the RMET in a US representative sample. I hypothesised that the RMET is a multidimensional measure of ToM ability and conducted EFA to identify the hypothesised multifactorial structure (H1a). Consistent with previous research, I failed to identify an appropriate factor structure for the RMET. Based on fit

indices, the best statistical model fit was obtained by retaining 12 factors, however, this model was not conceptually viable. Conversely, a three-factor model provided conceptual explanatory power with “flirtatious,” “thoughtful,” and “negative” factors, but resulted in poor statistical model fit. Together, the poor model fit, the failure of a subset of items to load as expected, and a high proportion of items failing to load on to any factor, suggest that, despite making conceptual sense, the three-factor model should be rejected.

I evaluated the possibility that conducting separate analyses on the data from participants with high versus low levels of autistic traits would result in separate factor structures and better model fit for both groups (H1b). While I did find evidence that within my sample the factor structure was different for the two groups, model fit did not increase for either group. Rather, the fit indices indicated worse fit for both groups and many items still failed to load on to any factor.

I also considered the possibility that gaze direction could differentially influence the RMET performance of individuals with higher levels of autistic traits. There was a mix of direct and indirect gaze direction in all three of the factors identified in the group with high AQ-28 scores, indicating that gaze direction was not an underlying factor for the performance of individuals with higher levels of autistic traits.

4.2. Validity: Does the RMET Rely on Mental State Reasoning? (H2a-H2c)

A second primary aim of this study was to evaluate whether the RMET relies on mental state reasoning in addition to emotion recognition. I hypothesised that both the AQ-28 and TAS would negatively correlate with RMET scores (H2a), but that the relationship between RMET scores and AQ-28 scores would no longer be significant after controlling for TAS scores (H2b), indicating that the RMET relies on emotion recognition but not mental state reasoning. Contrary to my prediction, total AQ-28 scores did not correlate with RMET scores, even without controlling for TAS scores. The relationship between RMET scores and total TAS and AQ-28 scores within my sample was consistent with Oakley et al.’s (2016) finding that TAS scores are more predictive of RMET performance than

Autism Spectrum Quotient scores. However, looking at the subscales of the AQ-28 revealed a more complicated relationship.

Whereas total AQ-28 scores did not correlate with RMET scores, two of the AQ-28 subscales, *imagination* and *social skills*, did, but in opposite directions. The *social skills* subscale positively correlated with RMET scores, with RMET scores increasing with an increase in autistic traits related to *social skills*. Examination of the items in this subscale indicate that it is closely related to levels of comfort or enjoyment in social situations ("I find social situation easy", "I prefer to do things with others rather than on my own"). It does not necessarily follow that a person who prefers solitary tasks or feels uncomfortable in social situations has a deficit in ToM ability. One possible explanation for the positive correlation between the *social skills* subscale and RMET scores is that individuals who feel less comfortable in social interactions spend more time observing faces, resulting in an improved ability to read mental states from eyes.

In contrast, the *imagination* subscale correlated negatively with RMET scores. This subscale, which includes questions such as "I find it easy to work out what someone is thinking or feeling" and "I find it difficult to work out people's intentions," relates closely to ToM ability. The negative correlation between the *imagination* subscale and RMET scores does suggest that the RMET relies on mental state reasoning in addition to emotion recognition.

Exploratory hierarchical regression analysis suggested that emotion recognition as indexed by the TAS, and ToM as indexed by the AQ-28, had a similar impact on RMET scores. However, the best predictor of RMET scores in my sample was IMT memory scores. The IMT memory questions which could be interpreted as a measure of general cognitive ability, making these results consistent with previous research that has found correlations between RMET performance and other cognitive abilities including IQ and verbal memory (Baker et al., 2014; Dalkner et al., 2019). Notably, the combined influence of all variables in the hierarchical regression model only accounted for approximately 25% of the variance in RMET performance. This might indicate that the RMET taps additional capacities that are separate from those measured by the other tasks in this study, or as

discussed in section 4.5.2, the task may evoke highly idiosyncratic responses, making the test unsuitable for group level comparisons.

I also hypothesised that AQ-28, but not TAS scores would correlate with the IMT ToM scores (H2c). Although the AQ-28 *imagination* subscale and all three TAS subscales had significant correlations with IMT ToM scores, none of these relationships remained significant after controlling for the IMT memory scores.

4.3. Comfort Viewing Eye Stimuli and RMET Performance (H3)

As hypothesised, levels of comfort positively correlated with RMET scores (H3a): individuals who felt more comfortable viewing the eye stimuli performed better on the task. Also as hypothesised (H3b), AQ-28 total scores correlated negatively with comfort viewing eye stimuli. This is consistent with previous research showing that autistic individuals have a negative response to eye stimuli (Trevisan et al., 2017). However, these correlations do not indicate direction of causality, and it should be noted that my rating of comfort was based off a single self-report measure that participants completed immediately after the RMET. Research also shows anxiety is more prevalent in autistic populations than the general population (Hollocks et al., 2019), which could also contribute to lower reported levels of comfort. Participants who found the RMET more difficult might have rated their comfort viewing the stimuli lower due to the discomfort of the task rather than the eye stimuli per se. However, these results suggest that further assessment of the role of comfort viewing eyes on RMET performance is warranted. Because the RMET was originally validated by the performance of autistic individuals (Baron-Cohen, Wheelwright, Hill et al., 2001), if comfort viewing eye stimuli is a significant factor in the performance of autistic individuals, it will have implications for the use of the RMET in other clinical and nonclinical groups in which the nature of the stimuli may or may not be aversive.

While the mediation analysis of the *imagination* and *social skills* subscales both showed a significant mediating effect of comfort on the relationship between these AQ-28 subscales and the

RMET (H3c), in both cases the mediated effects were very small, with the upper end of the confidence interval close to zero (-.02).

4.4. CFA of the TAS and AQ-28 Subscales (H4 and H5)

Absolute fit indices indicated acceptable model fit for both the TAS and the first-order AQ-28 subscales, however, relative fit indices were below the guidelines for acceptable model fit. I did identify and remove participants who provided straight lined responses to the TAS and AQ-28, which improved model fit. Straight lining is just one type of random responding, and it is likely that some participants may have randomly responded without straight lining. Despite relative model fit indices that were slightly below the recommended cutoff for the TAS, my findings broadly support the factor structure of the TAS in a US representative sample when administered in an online format. I did find low levels of internal consistency for the *externally oriented thinking* subscale, consistent with levels reported by Bagby et al. (1994; 2014), suggesting that this subscale does not capture a single latent construct, which could also reduce overall model fit. This subscale should be interpreted with caution. For the AQ-28, the CFI in my study (.824) was marginally lower than the value of .86 reported in Hoekstra et al.'s (2011) validation study of the AQ-28. Internal consistency of the scale as a whole was acceptable, however, three of the five subscales did not reach acceptable levels of internal consistency. These results support the use of the full AQ-28 as an online measure in a US population, however, there is a need for further investigation of the validity of the *routine, imagination,* and *numbers* subscales.

The second-order AQ-28 factor structure resulted in less optimal model fit indices than the first-order model. This suggests that there is no value in adding the second order factors to distinguish between the *numbers and patterns* factor and a *social behaviour* factor that incorporates the other four subscales.

4.5. Theoretical Questions About the Validity of the RMET

Two theoretical questions emerged from my study that have important implications for the use of the RMET as a measure of social cognition: (1) Is the RMET valid across different populations? (2) What makes some RMET items more challenging than others?

4.5.1. Question 1: Is the RMET valid across different populations?

As noted above, the validity of the target mental states in the RMET is based on consensus. Baron-Cohen, Wheelwright, Hill et al. (2001) validated the individual RMET items in a combined sample of 225 participants consisting of 103 Cambridge University students and 122 members of the general public in Cambridge and Essex, UK. The cut-off levels for consensus were "arbitrarily selected but with the aim of checking that a clear majority of the normal controls selected the target word and that this was selected at least twice as often as any foil" (Baron-Cohen, Wheelwright, Hill, et. al., 2001, p. 244). In my US representative sample of 1,181 members of the general public, eight items (22%) failed to pass the original criteria for retention in the test (i.e. ≥ 50% of participants selecting the target and ≤ 25% of participants selecting the same incorrect foil). This raises the question: are the targets for these items valid in my sample?

Researchers validating translated versions of the RMET have concluded that the RMET is valid despite a subset of items failing to meet the original criteria for retention in their samples. In a validation study of the Italian translation of the RMET conducted by Vellante et al. (2013) in a sample of 200 university students, more than 25% of participants selected the same incorrect foil for seven items, two of which (10, 17) overlap with problematic items identified in my sample. Two of these items also had a target response rate below 50%. In a study validating a Korean translation of the RMET using images of Asian eyes, Lee and Nam (2020) reported 5 items for which less than 50% of participants selected the target. They did not report the percentage of participants selecting each foil. Prevost et al., (2014) validated a French version of the RMET by comparing results on the English (N= 109) and French (N = 97) versions of the test. For the French version of the test, less than 50% of participants selected the target for seven items, three of which also had a target selection rate

below 50% in the English sample. The authors advised scoring two of these items separately for the French version because they had “clearly worse scores than the English version” (Prevost et al., 2014, p. 199). An additional five items on the English version and three items on the French version had a foil that was selected by ≥ 25% of participants. Yet, the conclusion of the authors of all three of these validation studies was that the translated versions are valid. Despite these conclusions, if consensus remains the only source of validation for the target mental states then we cannot be sure that these targets are valid for a sample in which the criteria for consensus are not met.

One potential solution would be to edit or remove the problematic items. Previous researchers have created psychometrically driven abbreviated versions of the RMET, however, there has been little consistency in the test items that are retained across abbreviated versions (see Table 1). In addition to identifying problematic test items statistically, it is essential to consider why there is so much variation in the levels of consensus (and other psychometric properties of the RMET) across different samples and what the implications are for the RMET as a measure of social cognition. We need theoretical explanations for why the targets are selected less frequently for certain RMET items. In the absence of such theoretical explanations, it is unclear whether a reduced version based on performance in one sample would be suitable for use in another sample.

4.5.2. Question 2: What makes some RMET items more challenging than others?

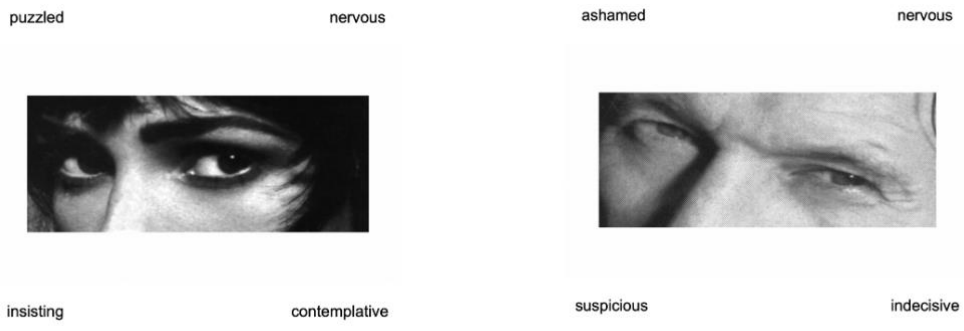
A key assumption of the RMET is that the ability to select the target over the foils indicates ToM ability, with the implicit underlying assumption that more difficult RMET items require higher levels of ToM ability. Part of the motivation for designing the RMET was to overcome ceiling effects in existing ToM tests, and the test is supposed to be challenging. However, Baron-Cohen, Wheelwright, Hill, et al. (2001) did not provide information on which test items should be more challenging or provide theoretical explanations for why some items should be more challenging. Where researchers have conducted analyses based on difficulty of RMET items, difficulty has been determined by performance on the test rather than theoretically derived. For example, Baltazar et al. (2020) used the percentage correct for each item published in Prevost et al. (2014) to determine

item difficulty, and Burke et al. (2020) calculated a difficulty coefficient for each item based on the ratio of overall performance on each item relative to the performance of the highest and lowest scoring participants. What remains unclear is what factors influence the difficulty of test items and whether these factors are linked to ToM ability.

It is possible that the assumption that more difficult RMET items place greater demands on ToM abilities is correct and that individual differences in test performance are the result of varying levels of ToM ability. This assumption is compatible with both a unidimensional and multidimensional factor structure. In the case of unidimensional structure, there would be a single underlying factor driving performance, whereas different facets of ToM ability could drive performance if the RMET is a multidimensional test. The current results do not support this assumption. Factor analyses failed to identify either a unidimensional or multidimensional factor structure for the RMET. Additionally, looking at the items with the highest and lowest correct response rate in my sample, there is no clear theoretical explanation for why one item should be more difficult than the other in terms of comparative demands on ToM ability (see Figure 5).

Figure 5

RMET Items with the Highest and Lowest Percentage of Correct Responses



Note. The image on the left is RMET item 35. Only 36% of participants in my sample selected the target response (nervous). The image on the right is RMET item 36. The target response (suspicious) was selected by 83% of participants in my sample. Image from the Autism Research Centre (2020) https://www .autismresearchcentre.com/arc_tests

An alternative explanation is that the difficulty of RMET items relates to the extent to which the target and foils match the images. Critically, a consensus rate above 50% is not equivalent to an endorsement that a target is a "good" fit for the image. Participants may be adopting a strategy of choosing the 'least bad' response. To my knowledge, no study has asked participants to rate *how well* the target and foil mental states match the images. The possibility exists that the more challenging RMET items are difficult because (1) more than one of the choices match the image or (2) none of the choices match the image. I consider these two possibilities below.

4.5.2.1. Difficulty Increases when the Target is not Inherently More Correct than the Foils.

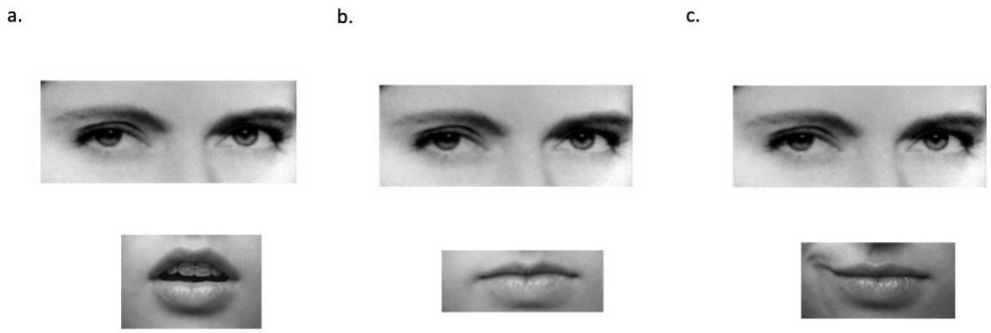
Some participants in my study commented that determining what a person is thinking or feeling from eyes without the whole face or any gestures is difficult. Removing the eyes from the context of the face does likely contribute to the difficulty of the RMET. It could also potentially explain why some test items are more difficult than others. There is evidence that participants rely more on information from the eyes for some emotions (sadness and fear) and the mouth for others (disgust and happiness, Wegrzyn et al., 2017). Items with target mental states for which participants tend to rely more heavily on the configuration of the mouth than the eyes might be more difficult when no mouth is present. However, if this is the case, it would be critical to ensure that there is sufficient information in the eyes to identify the target as the most appropriate mental state. If there is insufficient information in the image to differentiate the target from the foils, then we cannot be sure what information participants are basing their choices on. They may be filling in the gaps with their own contextual information that makes the choice of a foil valid.

RMET item 17 failed to pass the original criteria for retention in my study and multiple other studies (e.g. Olderbak et al., 2015; Prevost et al., 2014 Vellante et al., 2013). This suggests that there may be insufficient information in the image to guide the selection of the target over the foils. Face validity for this item is poor: it is unclear what information in the image is supposed to indicate that the answer is "doubtful" rather than "affectionate," "playful," or "aghast." It is possible that, given appropriate contextual information, the image is consistent with more than one of the choices.

Figure 6 shows the eyes from item 17 with additional face cues to demonstrate how different contextual information might impact which mental state term best matches the image. Combined with mouth (a), the eyes might be "aghast." Mouth (c) makes the eyes look "affectionate" or "playful," whereas mouth (b) could be consistent with a "doubtful" expression. In the absence of sufficient information in the image itself to definitively indicate the correct response, we cannot evaluate the validity of participants' responses because we cannot be sure what their responses are based on. It is even conceivable that individuals with higher levels of ToM ability are better able to adaptively match multiple mental state terms to the image, making item 17 more difficult for people with higher levels of ToM ability, which would be consistent with our finding that higher AQ-28 social skills scores, which indicate difficulty with social skills, positively correlated with RMET scores.

Figure 6

RMET Item 17 Combined with Different Mouth Expressions



Note. (a) mouth from a surprised expression (b) mouth from a neutral expression (c) mouth with a cheeky grin. The target for item 17 is "doubtful." The three foils are "affectionate," "playful," and "aghast." In my sample, 53% of respondents selected the target and 26% selected "affectionate." The expression in the eyes is relatively neutral and adding a mouth as contextual information shows that the eyes could be consistent with a variety of mental states. Eye images from the RMET from Autism Research Centre (2020) https://www .autismresearchcentre.com/arc_tests. Mouth images adapted from Piacquadio (2020).

4.5.2.2. Difficulty Increases when the Target does not Match the Image.

Some participants left comments indicating that they did not think any of the choices were appropriate for some of the RMET items, with one suggesting I add a "none of the above" option. As an example, the target for item 34 (see Figure 1) is "distrustful" and the foils are "terrified" "baffled" and "aghast." In my sample, 57% of participants selected the target and 27% selected "baffled." Scott et al. (2011) had forty university students rate the valence of the RMET images when presented without the mental state terms. The image from item 34 was rated as having positive valence, which is inconsistent with the target and the foils. In this case, the difficulty of the item might come from a mismatch between information in the image and the mental state choices with none of the choices being a good match for the image.

4.6. Implications for Future Use of the RMET

I did not find evidence for a unidimensional or multidimensional factor structure underlying the RMET. Consistent with previous studies, the psychometric properties of the test as a whole and of individual items were lacking. Most problematically, a significant proportion of items failed to meet the validity criterion established by the test creators. Considering these issues, I suggest that the RMET should not be used as a measure of social cognition. Currently, it is not clear that differences in performance on this test relate to differences in mental state reasoning, emotion recognition, or other aspects of social cognition, which makes it challenging to meaningfully interpret performance on the test.

The widespread use of the RMET despite multiple reports highlighting psychometric shortcomings of the test is puzzling and may indicate insufficient attention being paid to measurement validity. Flake and Fried (2019, p. 8) identify a number of "questionable measurement practices" in psychological research, which they define as "decisions researchers make that raise doubts about the validity of measure use in a study and ultimately the final conclusion." Flake and Fried recommend three steps to avoid questionable measurement practices: (1) provide theoretical definitions of the psychological construct under investigation (2) justify selected measures by

explicitly stating how they target the construct(s) under investigation and (3) provide evidence for the validity of the measure(s). I contend that the RMET would not be selected as a valid measure of ToM by researchers who take these steps in their research. As noted in the introduction, ToM is a difficult construct to define. However, researchers must still operationalise the construct in some way and explain how their selected measures capture ToM as they define it. After doing so, it would be difficult, if not impossible, to justify the selection of the RMET as a measure of ToM because, in order to do so, researchers would need to understand how the RMET targets ToM ability. The current findings, and those from multiple previous validation studies (e.g. Prevost et al., 2014; Vellante et al., 2013) have failed to demonstrate that the RMET *does* measure ToM ability, let alone *how*.

4.7. Limitations

The main limitation of this study was that I did not include any autistic participants. Differences in RMET abilities were evaluated on autistic traits in a non-clinical population. It is possible that a factor structure would emerge in a clinical population. In this case, it might be possible to argue that the RMET is valid for use in clinical populations. However, the theoretical issues identified with the RMET would still exist.

Another potential limitation of this study is data quality. Data was collected from participants completing surveys online for compensation. An attention check question was included to identify insincere responders. Considering a large number of participants failed the attention check and a subset of participants who passed the attention check provided straight lined responses, despite our best efforts, it is likely that the data that was analysed still contained some random or insincere responses. However, I believe this may be the largest demographically representative sample to have been used in research on the RMET, and this should mitigate the influence of any remaining insincere responses.

The IMT was always presented last instead of being counterbalanced with the RMET, AQ-28, and TAS. This decision was made because the IMT had the most demanding reading comprehension

and it might have led to participant fatigue or disengagement with subsequent tasks. However, this could have introduced order effects, and participants may have been fatigued before completing the IMT, thus reducing performance on this measure. A number of participants commented that they had difficulty remembering the names of the characters in the IMT story. They were not able to refer to the story when answer the questions. While the questions do require ToM knowledge to answer correctly, this task may have relied too heavily on memory, thus reducing its ability to capture differences in ToM ability.

Additionally, while my sample was demographically representative, the mean AQ score was higher than expected in a random sample. As noted above, this could relate to the characteristics of individuals who complete surveys online for money. However, despite the higher mean score, the highest scores in the low AQ trait group (62) were still below the cutoff level identified by Hoekstra et al. as indicative of autism (≥ 65).

As noted in the results section, the mean RMET score in my sample was lower than that reported by Baron-Cohen et al. (2001) and Olderbak et al. (2015). This could be related to the high mean scores on the AQ-28, since it has been found that individuals with more autistic traits perform worse on the RMET (Gökçen et al., 2016). However, if that were the case, I would expect to find a correlation between AQ-28 total scores and RMET scores, which I did not. Another possibility is that the low scores relate the theoretical issues with the RMET identified in the discussion. Of note, lower mean scores in nonclinical populations were also reported in validation studies of translated versions of the RMET (e.g. Khorashad et al., 2015 [22.76], Pfaltz et al., 2013 [24.5], Vellante et al., 2013 [24.9]).

4.8. Conclusion

The RMET is a widely used measure of ToM ability in a variety of clinical and nonclinical populations. Despite being perceived as a well-validated tool, this study raises considerable doubts about the validity of the RMET as a measure of social cognition. Researchers should stop using the RMET unless a satisfactory theoretical explanation of how it measures ToM can be proposed and

tested. Researchers should also avoid citing conclusions based on this measure as it is not possible to interpret the significance of performance on the RMET in absence of a theoretical understanding of what it is measuring.

References

Abell, F., Happé, F., & Frith, U. (2000). Do triangles play tricks? Attribution of mental states to animated shapes in normal and abnormal development. *Cognitive Development, 15*(1), 1-16. https://doi.org/10.1016/s0885-2014(00)00014-9

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

Ahmed, F., & Stephen Miller, L. (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

Autism Research Centre (2020). Eyes Test (adult). Retrieved from https://www.autismresearchcentre.com/tests/eyes-test-adult/

Bado, F. M. R., Rebustini, F., Jamieson, L., Cortellazzi, K. L., & Mialhe, F. L. (2018). Evaluation of the psychometric properties of the Brazilian version of the Oral Health Literacy Assessment in Spanish and development of a shortened form of the instrument. *PloS One, 13*(11), e0207989-e0207989. https://doi.org/10.1371/journal.pone.0207989

Bagby, R. M., Ayearst, L. E., Morariu, R. A., Watters, C., & Taylor, G. J. (2014). The internet administration version of the 20-Item Toronto Alexithymia Scale. *Psychological Assessment*, *26*(1), 16–22. https://doi.org/10.1037/a0034316

Bagby, R. M., Parker, J. D. ., & Taylor, G. J. (1994). The twenty-item Toronto Alexithymia

scale—I. Item selection and cross-validation of the factor structure. *Journal of Psychosomatic Research*, *38*(1), 23–32. https://doi.org/10.1016/0022-3999(94)90005-1

Baker, C. A., Peterson, E., Pulos, S., & Kirkland, R. A. (2014). Eyes and IQ: A meta-analysis of the relationship between intelligence and "Reading the Mind in the Eyes". *Intelligence (Norwood), 44*(1), 78-92. https://doi.org/10.1016/j.intell.2014.03.001

Baltazar, M., Geoffray, M. M., Chatham, C., Bouvard, M., Martinez Teruel, A., Monnet, D., Scheid, I., Murzi, E., Couffin-Cadiergues, S., Umbricht, D., Murtagh, L., Delorme, R., Le-Moal, M. L., Leboyer, M., & Amestoy, A. (2020). "Reading the Mind in the Eyes" in autistic adults is modulated by valence and difficulty: An InFoR Study. *Autism Research*. https://doi.org/10.1002/aur.2390

Baron-Cohen, S. (1995). *Learning, development, and conceptual change. Mindblindness: An essay on autism and theory of mind.* Cambridge, MA, US: The MIT Press.

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-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., & Wheelwright, S. (2004). The Empathy Quotient: An investigation of adults with Asperger Syndrome or High Functioning Autism, and normal sex differences. *Journal of Autism and Developmental Disorders, 34*(2), 163-175. https://doi.org/10.1023/b:jadd.0000022607.19833.00

Baron-Cohen, S., Wheelwright, S., Hill, J., Raste, Y., & Plumb, I. (2001). The "Reading the

Mind in the Eyes" Test revised version: a study with normal adults, and adults with Asperger syndrome or high-functioning autism. *Journal of Child Psychology and Psychiatry, and Allied Disciplines*, *42*(2), 241–251. http://dx.doi.org/10.1111/1469-7610.00715

Baron-Cohen, S., Wheelwright, S., Skinner, R., Martin, J., & Clubley, E. (2001). The Autism Spectrum Quotient (AQ): Evidence from Asperger syndrome/high-functioning autism, males and females, scientists and mathematicians. *Journal of Autism and Developmental Disorders, 31*(1), 5-17. https://doi.org/10.1023/A:1005653411471

Bergh, D. (2015). Sample size and chi-squared test of fit—A comparison between a random sample approach and a chi-square value adjustment method using Swedish adolescent data. In *Pacific Rim Objective Measurement Symposium (PROMS) 2014 Conference Proceedings* (pp. 197–211). Berlin, Heidelberg: Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-662-47490-7_15

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

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

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

Burke, T., Pinto-Grau, M., Costello, E., Peelo, C., Lonergan, K., Heverin, M., Hardiman, O., & Pender, N. (2020). The reading the mind in the eyes test short form (A & B): validation and outcomes in an amyotrophic lateral sclerosis cohort. *Amyotrophic Lateral Sclerosis and Frontotemporal Degeneration, 21*(5-6), 380-388. https://doi.org/10.1080/21678421.2020.1772824

Burke, T., Pinto-Grau, M., Lonergan, K., Elamin, M., Bede, P., Costello, E., Hardiman, O., & Pender, N.

(2016). Measurement of social cognition in Amyotrophic Lateral Sclerosis: A population based study. *PloS One*, *11*(8), e0160850–e0160850. https://doi.org/10.1371/journal.pone.0160850

Charernboon, T., & Lerthattasilp, T. (2017). The Reading the Mind in the Eyes test: Validity and reliability of the Thai version. *Cognitive and Behavioral Neurology, 30*(3), 98-101. https://doi.org/10.1097/WNN.0000000000000130

Christmann, A., & Van Aelst, S. (2006). Robust estimation of Cronbach's alpha. *Journal of Multivariate Analysis, 97*(7), 1660-1674. https://doi.org/10.1016/j.jmva.2005.05.012

Clark, L. A., & Watson, D. (1995). Constructing validity: Basic issues in objective scale development. *Psychological Assessment, 7*(3), 309-319. https://doi.org/10.1037/1040-3590.7.3.309

Coppock, A., & McClellan, O. A. (2019). Validating the demographic, political, psychological, and experimental results obtained from a new source of online survey respondents. *Research & politics, 6*(1). 205316801882217. https://doi.org/10.1177/2053168018822174

Dalkner, N., Bengesser, S. A., Birner, A., Fellendorf, F. T., Hamm, C., Platzer, M., Oilz, R., Queissner, R., Rieger, A., Weber, B., Kapfhammer, H. P., Weiss, E. M., & Reininghaus, E. Z. (2019). The relationship between "Eyes Reading" ability and verbal memory in bipolar disorder. *Psychiatry Research, 273*, 42-51. https://doi.org/10.1016/j.psychres.2019.01.015

Dordevic, J., Zivanovic, M., Pavlovic, A., Mihajlovic, G., Karlicic, I., & Pavlovic, D. (2017). Psychometric evaluation and validation of the Serbian version of “Reading the Mind in the Eyes” test. *Psihologija, 50*(4), 483-502. https://doi.org/10.2298/PSI170504010D

Dunn, T. J., Baguley, T., & Brunsden, V. (2014). From alpha to omega: A practical solution to the pervasive problem of internal consistency estimation. *The British journal of psychology, 105*(3), 399-412. https://doi.org/10.1111/bjop.12046

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. M. (2019). What do you have in mind? Measures to assess mental state reasoning in neuropsychiatric populations. *Frontiers in psychiatry, 10*. https://doi.org/10.3389/fpsyt.2019.00425

Espinós, U., Fernández-Abascal, E. G., & Ovejero, M. (2018). What your eyes tell me: Theory of mind in bipolar disorder. *Psychiatry Research, 262*, 536-541. https://doi.org/10.1016/j.psychres.2017.09.039

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

Fertuck, E. A., Jekal, A., Song, I., Wyman, B., Morris, M. C., Wilson, S. T., Brodsky, B.S., & Stanley, B. (2009). Enhanced 'Reading the Mind in the Eyes' in borderline personality disorder compared to healthy controls. *Psychological Medicine, 39*(12), 1979-1988. https://doi.org/10.1017/S003329170900600X

Flake, J. K., & Fried, E. I. (2019). *Measurement schmeasurement: Questionable measurement practices and how to avoid them.* PsyArXiv. https://doi.org/10.31234/osf.io/hs7wm

Flora, D. B. (2020). Your Coefficient Alpha Is Probably Wrong, but Which Coefficient Omega Is Right? A Tutorial on using R to obtain better reliability estimates. *Advances in Methods and Practices in Psychological Science*. https://doi.org/10.1177/2515245920951747

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 and Psychotherapy, 24*(6), 1421-1434. https://doi.org/10.1002/cpp.2091

Franklin, R. G., & Zebrowitz, L. A. (2016). Aging-related changes in decoding negative

complex mental states from faces. *Experimental Aging Research*, *42*(5), 471–478. https://doi.org/10.1080/0361073X.2016.1224667

Furr, R. (2011). *Scale construction and psychometrics for social and personality psychology*. SAGE.

Gernsbacher, M. A., & Yergeau, M. (2019). Empirical failures of the claim that autistic people lack a theory of mind. *Archives of Scientific Psychology, 7*(1), 102-118. https://doi.org/10.1037/arc0000067

Giordano, M., Licea-Haquet, G., Navarrete, E., Valles-Capetillo, E., Lizcano-Cortés, F., Carrillo-Peña, A., & Zamora-Ursulo, A. (2019). Comparison between the Short Story Task and the Reading the Mind in the Eyes Test for evaluating Theory of Mind: A replication report. *Cogent Psychology, 6*(1). https://doi.org/10.1080/23311908.2019.1634326

Girli, A. (2014). Psychometric Properties of the Turkish child and adult form of "Reading the Mind in the Eyes Test." *Psychology (Irvine, Calif.)*, *5*(11), 1321–1337. https://doi.org/10.4236/psych.2014.511143

Gökçen, E., Frederickson, N., & Petrides, K. (2016). Theory of mind and executive control deficits in typically developing adults and adolescents with high levels of autism traits. *Journal of Autism and Developmental Disorders*, *46*(6), 2072–2087. https://doi.org/10.1007/s10803-016-2735-3

Goodboy, A. K., & Martin, M. M. (2020). Omega over alpha for reliability estimation of unidimensional communication measures. *Annals of the International Communication Association, 44*(4), 422-439. https://doi.org/10.1080/23808985.2020.1846135

Green, S. B., & Yang, Y. (2015). Evaluation of dimensionality in the assessment of internal consistency reliability: Coefficient alpha and omega coefficients. *Educational Measurement, Issues and Practice, 34*(4), 14-20. https://doi.org/10.1111/emip.12100

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

Hadjikhani, N., Åsberg Johnels, J., Zürcher, N. R., Lassalle, A., Guillon, Q., Hippolyte, L., Billstedt, E., Ward, N., Lemonnier, E., & Gillberg, C. (2017). Look me in the eyes: constraining gaze in the eye-region provokes abnormally high subcortical activation in autism. *Scientific Reports, 7*(1), 3163-3167. https://doi.org/10.1038/s41598-017-03378-5

Happé, F. (1994). An advanced test of theory of mind: Understanding of story characters' thoughts and feelings by able autistic, mentally handicapped, and normal children and adults. *Journal of Autism and Developmental Disorders, 24*(2), 129-154. https://doi.org/10.1007/bf02172093

Harkness, K. L., Jacobson, J. A., Duong, D., & Sabbagh, M. A. (2010). Mental state decoding in past major depression: Effect of sad versus happy mood induction. *Cognition and Emotion, 24*(3), 497-513. https://doi.org/10.1080/02699930902750249

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

Harrison, A., Tchanturia, K., & Treasure, J. (2010). Attentional bias, emotion recognition, and emotion regulation in anorexia: State or trait? *Biological Psychiatry (1969), 68*(8), 755-761. https://doi.org/10.1016/j.biopsych.2010.04.037

Hayton, J., Allen, D., & Scarpello, V. (2004). Factor retention decisions in exploratory factor analysis: A tutorial on parallel analysis. *Organizational Research Methods, 7*(2), 191-205. https://doi.org/10.1177/1094428104263675

Hoekstra, R., Vinkhuyzen, A., Wheelwright, S., Bartels, M., Boomsma, D., Baron-Cohen, S., Posthuma,

D., & Sluis, S. (2011). The construction and validation of an abridged version of the Autism-Spectrum Quotient (AQ-Short). *Journal of Autism and Developmental Disorders, 41*(5), 589-596. https://doi.org/10.1007/s10803-010-1073-0

Hollocks, M. J., Lerh, J. W., Magiati, I., Meiser-Stedman, R., & Brugha, T. S. (2019). Anxiety and depression in adults with autism spectrum disorder: a systematic review and meta-analysis. *Psychological Medicine, 49*(4), 559-572. https://doi.org/10.1017/S0033291718002283

Hudson, C. C., Shamblaw, A. L., Harkness, K. L., & Sabbagh, M. A. (2020). Valence in the Reading the Mind in the Eyes task. *Psychological Assessment*, *32*(7), 623–634. https://doi.org/10.1037/pas0000818

Jankowiak-Siuda, K., Baron-Cohen, S., Bialaszek, W., Dopierala, A., Kozlowska, A., & Rymarczyk, K. (2016). Psychometric evaluation of the 'reading the mind in the eyes' test with samples of different ages from a polish population. *Studia Psychologica: Journal for Basic Research in Psychological Sciences*, *58*(1), 18–31. https://doi.org/10.21909/sp.2016.01.704

Jones, C. R. G., Simonoff, E., Baird, G., Pickles, A., Marsden, A. J. S., 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

Jorgensen, T. D., Pornprasertmanit, S., Schoemann, A. M., & Rosseel, Y. (2020). semTools: Useful tools for structural equation modeling. R package version 0.5-3. Retrieved from https://CRAN.R-project.org/package=semTools

Khorashad, B., Baron-Cohen, S., Roshan, S., Kazemian, G., Khazai, M., Aghili, M., Talaei, L., & Afkhamizadeh, Z. (2015). The "Reading the Mind in the Eyes" test: Investigation of psychometric properties and test–retest reliability of the Persian version. *Journal of Autism and Developmental Disorders*, *45*(9), 2651–2666. https://doi.org/10.1007/s10803-015-2427-4

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

Kotrla Topić, M., & Perković Kovačević, M. (2019). Croatian Adaptation of the Revised Reading the Mind in the Eyes Test (RMET). *Psihologijske Teme, 28*(2), 377-395. https://doi.org/10.31820/pt.28.2.8

Kung, K.T.F. (2020). Autistic traits, systemising, empathising, and theory of mind in transgender and non-binary adults. *Molecular Autism, 11***(**1), 73. https://doi.org/10.1186/s13229-020-00378-7

Kylliäinen, A., & Hietanen, J. (2006). Skin conductance responses to another person's gaze in children with autism. *Journal of Autism and Developmental Disorders, 36*(4), 517-525. https://doi.org/10.1007/s10803-006-0091-4

Launay, J., Pearce, E., Wlodarski, R., van Duijn, M., Carney, J., & Dunbar, R. I. M. (2015). Higher-order mentalising and executive functioning. *Personality and Individual Differences, 86*, 6-14. https://doi.org/10.1016/j.paid.2015.05.021

Lawrence, E. J., Shaw, P., Baker, D., Baron-Cohen, S., & David, A. S. (2004) Measuring empathy: Reliability and validity of empathy quotient, *Psychological Medicine, 34*, 911–924. http://dx.doi.org/10.1017/S0033291703001624

Lee, H.-S., Corbera, S., Poltorak, A., Park, K., Assaf, M., Bell, M. D., Wexier, B. E., Cho, Y.-I., Jung, S., Brocke, S., & Choi, K.-H. (2018). Measuring theory of mind in schizophrenia research: Cross-cultural validation. *Schizophrenia Research, 201*, 187-195. https://doi.org/10.1016/j.schres.2018.06.022

Lee, H.-R., & Nam, G. (2020). Development and validation of the Korean version of the Reading the Mind in the Eyes Test. *PloS One, 15*(8), e0238309–e0238309. https://doi.org/10.1371/journal.pone.0238309

Li, T.-S., Liu, C.-M., Liu, C.-C., Hsieh, M. H., Lin, Y.-T., Wang, E.-N., Huang, T.-J., Chou, T.-L.

2020). Social cognition in schizophrenia: A network-based approach to a Taiwanese version of the Reading the Mind in the Eyes test. *Journal of the Formosan Medical Association, 119*(1), 439-448. https://doi.org/10.1016/j.jfma.2019.08.008

Mar, R. A., Oatley, K., Hirsh, J., Dela Paz, J., & Peterson, J. B. (2006). Bookworms versus nerds: Exposure to fiction versus non-fiction, divergent associations with social ability, and the simulation of fictional social worlds. *Journal of Research in Personality, 40*(5), 694-712. https://doi.org/10.1016/j.jrp.2005.08.002

Marsh, H. W., Hau, K.-T., & Wen, Z. (2004). In search of golden rules: Comment on hypothesis-testing approaches to setting cutoff values for fit indexes and dangers in overgeneralizing Hu and Bentler's (1999) findings. *Structural Equation Modeling, 11*(3), 320-341. https://doi.org/10.1207/s15328007sem1103_2

Marsh, P., Polito, V., Singh, S., Coltheart, M., Langdon, R., & Harris, A. (2016). A quasi-randomized feasibility pilot study of specific treatments to improve emotion recognition and mental-state reasoning impairments in schizophrenia. *BMC Psychiatry*, *16*(1), 360. https://doi.org/10.1186/s12888-016-1064-6

Meyer, J., & Shean, G. (2006). Social-cognitive functioning and schizotypal characteristics. *The Journal of Psychology, 140*(3), 199-207. https://doi.org/10.3200/JRLP.140.3.199-207

Mısır, E., Bora, E., & Akdede, B. B. (2018). Relationship between social-cognitive and social perceptual aspects of theory of mind and neurocognitive deficits, insight level and schizotypal traits in obsessive-compulsive disorder. *Comprehensive Psychiatry, 83*, 1-6. https://doi.org/10.1016/j.comppsych.2018.02.008

Müller, C. M., & Gmünder, L. (2014). An evaluation of the “Reading the Mind in the Eyes Test” With seventh to ninth graders. *Journal of Mental Health Research in Intellectual Disabilities, 7*(1), 34-44. https://doi.org/10.1080/19315864.2012.714055

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

Emotion. *Journal of Abnormal Psychology*, *125*, 1–25. http://dx.doi.org/10.1037/abn0000182

Olderbak, S., Wilhelm, O., Olaru, G., Geiger, M., Brenneman, M. W., & Roberts, R. D. (2015). A psychometric analysis of the reading the mind in the eyes test: Toward a brief form for research and applied settings. *Frontiers in Psychology*, *6*, 1–14. https://doi.org/10.3389/fpsyg.2015.01503

Öztürk, Y., Özyurt, G., Turan, S., Mutlu, C., Tufan, A. E., & Pekcanlar Akay, A. (2020). Association of theory of mind and empathy abilities in adolescents with social anxiety disorder. *Current Psychology.* 1-10. *https://doi.org/10.1007/s12144-020-00707-2*

Pahnke, R., Mau-Moeller, A., Hamm, A. O., & Lischke, A. (2020). Reading the Mind in the Eyes of Children Test (RME-C-T): Development and validation of a complex emotion recognition test. *Frontiers in Psychiatry, 11*, 376-376. https://doi.org/10.3389/fpsyt.2020.00376

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

Pfaltz, M. C., McAleese, S., Saladin, A., Meyer, A. H., Stoecklin, M., Opwis, K., Damman, G., & Martin Soelch, C. (2013). The reading the mind in the eyes test: test-retest reliability and preliminary Rsychometric Rroperties of the German version. *International Journal of Advances in Psychology, 2*(1), 1. https://doi.org/10.5167/uzh-87335

Piacquadio, A. (2020). *Collage photo of woman* [photograph]. Pexel. https://www.pexels.com/photo/collage-photo-of-woman-3812743/

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

Preti, A., Vellante, M., & Petretto, D. (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

Prevost, M., Carrier, M.-E., Chowne, G., Zelkowitz, P., Joseph, L., & Gold, I. (2014). The Reading the Mind in the Eyes test: validation of a French version and exploration of cultural variations in a multi-ethnic city. *Cognitive Neuropsychiatry, 19*(3), 189-204. https://doi.org/10.1080/13546805.2013.823859

R Core Team (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Raffo De Ferrari, A., Lagravinese, G., Pelosin, E., Pardini, M., Serrati, C., Abbruzzese, G., & Avanzino, L. (2015). Freezing of gait and affective theory of mind in Parkinson disease. *Parkinsonism & Related Disorders*, *21*(5), 509–513. https://doi.org/10.1016/j.parkreldis.2015.02.023

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

Revelle W (2020). *Psych: Procedures for Psychological, Psychometric, and Personality Research*. Northwestern University, Evanston, Illinois. R package version 2.0.7, https://CRAN.R-project.org/package=psych.

Richard-Mornas, A., Mazzietti, A., Koenig, O., Borg, C., Convers, P., & Thomas-Antérion, C. (2014). Emergence of hyper empathy after right amygdalohippocampectomy. *Neurocase, 20*(6), 666-670. https://doi.org/10.1080/13554794.2013.826695

Rominger, C., Bleier, A., Fitz, W., Marksteiner, J., Fink, A., Papousek, I., & Weiss, E. M. (2016). Auditory top-down control and affective theory of mind in schizophrenia with and without hallucinations. *Schizophrenia Research, 174*(1-3), 192-196. https://doi.org/10.1016/j.schres.2016.05.006

Rosseel Y. (2012). "lavaan: An R Package for Structural Equation Modeling." *Journal of Statistical Software*, *48*(2), 1–36. http://www.jstatsoft.org/v48/i02/. \

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

Ruzich, E., Allison, C., Smith, P., Watson, P., Auyeung, B., Ring, H., & Baron-Cohen, S. (2015). Measuring autistic traits in the general population: a systematic review of the Autism-Spectrum Quotient (AQ) in a nonclinical population sample of 6,900 typical adult males and females. *Molecular Autism*, *6*(1), 2. https://doi.org/10.1186/2040-2392-6-2

Sadeghi Bahmani, D., Razazian, N., Motl, R. W., Farnia, V., Alikhani, M., Pühse, U., Gerber, M., & Brand, S. (2020). Physical activity interventions can improve emotion regulation and dimensions of empathy in persons with multiple sclerosis: An exploratory study. *Multiple Sclerosis and Related Disorders*, *37*, 101380–101380. https://doi.org/10.1016/j.msard.2019.101380

Schmitt, H. S., Sindermann, C., Li, M., Ma, Y., Kendrick, K. M., Becker, B., & Montag, C. (2020). The dark side of emotion recognition – Evidence from cross-cultural research in Germany and China. *Frontiers in Psychology, 11*, 1132–1132. https://doi.org/10.3389/fpsyg.2020.01132

Scott, L. N., Levy, K. N., Adams, R. B., & Stevenson, M. T. (2011). Mental state decoding abilities in young adults with borderline personality disorder traits. *Personality Disorders: Theory, Research, and Treatment, 2*(2), 98-112. https://doi.org/10.1037/a0020011

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

Shamay-Tsoory, S. G., & Aharon-Peretz, J. (2007). Dissociable prefrontal networks for cognitive and affective theory of mind: A lesion study. *Neuropsychologia, 45*(13), 3054-3067. https://doi.org/10.1016/j.neuropsychologia.2007.05.021

Stiller, J., & Dunbar, R. I. M. (2007). Perspective-taking and memory capacity predict social network size. *Social networks, 29*(1), 93-104. https://doi.org/10.1016/j.socnet.2006.04.001

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

Sullivan, K., Zaitchik, D., & Tager-Flusberg, H. (1994). Preschoolers can attribute second-order beliefs. *Developmental Psychology, 30*(3), 395–402. https://doi.org/10.1037/0012-1649.30.3.395

Susana, L., Adoración, F., Ana, H., & Inés, T. (2017). The exploratory factor analysis of items: guided analysis based on empirical data and software. *Anales de Psicología, 33*(2), 417-432. https://doi.org/10.6018/analesps.33.2.270211

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

Tayfun, K., & Semra, Y. (2019). Theory of mind and related factors in parents of children diagnosed with autism spectrum disorders. *Klinik Psikiyatri Dergisi, 22*(2), 139-147. https://doi.org/10.5505/kpd.2018.83007

Tingley D., Yamamoto T., Hirose K., Keele L., Imai K. (2014). "mediation: R Package for Causal Mediation Analysis." *Journal of Statistical Software*, **59**(5), 1–38. http://www.jstatsoft.org/v59/i05/.

Trevisan, D. A., Roberts, N., Lin, C., & Birmingham, E. (2017). How do adults and teens with self-declared Autism Spectrum Disorder experience eye contact? A qualitative analysis of first-hand accounts. *PloS One, 12*(11), e0188446–e0188446. https://doi.org/10.1371/journal.pone.0188446

Vellante, M., Baron-Cohen, S., Melis, M., Marrone, M., Petretto, D. R., Masala, C., & Preti, A. (2013). The "Reading the Mind in the Eyes" test: Systematic review of psychometric properties and a validation study in Italy. *Cognitive Neuropsychiatry, 18*(4), 326-354. https://doi.org/10.1080/13546805.2012.721728

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

Warrier, V., Grasby, K. L., Uzefovsky, F., Toro, R., Smith, P., Chakrabarti, B., Khadake, J., Mawbey-Adamson, E., Litterman, N., & Hottenga, J.-J. (2018). Genome-wide meta-analysis of cognitive empathy: heritability, and correlates with sex, neuropsychiatric conditions and cognition. *Molecular Psychiatry, 23*(6), 1402-1409. https://doi.org/10.1038/mp.2017.122

Washburn, D., Wilson, G., Roes, M., Rnic, K., & Harkness, K. L. (2016). Theory of mind in social anxiety disorder, depression, and comorbid conditions. *Journal of Anxiety Disorders*, *37*, 71–77. https://doi.org/10.1016/j.janxdis.2015.11.004

Wegrzyn, M., Vogt, M., Kireclioglu, B., Schneider, J., & Kissler, J. (2017). Mapping the emotional face. How individual face parts contribute to successful emotion recognition. *PloS One, 12*(5), e0177239-e0177239. https://doi.org/10.1371/journal.pone.0177239

Youssef, F. F., Nunes, P., Sa, B., & Williams, S. (2014). An exploration of changes in cognitive and emotional empathy among medical students in the Caribbean. *International Journal of Medical Education, 5*, 185-192. https://doi.org/10.5116/ijme.5412.e641

Zinbarg, R. E., Yovel, I., Revelle, W., & McDonald, R. P. (2006). Estimating generalizability to a latent variable common to all of a scale's indicators: A comparison of estimators for ωh. *Applied Psychological Measurement, 30*(2), 121-144. https://doi.org/10.1177/0146621605278814

Appendix A

IMT Materials

Adapted from Launay, J., Pearce, E., Wlodarski, R., van Duijn, M., Carney, J., & Dunbar, R. I. M. (2015). Higher-order mentalising and executive functioning. *Personality and Individual Differences, 86*, 6-14. https://doi.org/10.1016/j.paid.2015.05.021

Below is a brief story. Please read the story carefully. We are going to ask you some questions about it.

The Cafeteria

Hannah was a bit late in getting to the cafeteria and by the time she got her lunch, there weren't many seats left. She noticed that there was one place free at the table where Emma and her boyfriend Matt and their friends always sat. So she went over and asked them if she could sit with them. Emma looked up and said "Oh actually, I was saving that seat for Abbie. Sorry!" So Hannah kept walking around the cafeteria trying to find somewhere to sit.

Eventually she sat down with her colleague Carolyn at a table in the corner. Carolyn noticed that Hannah was looking upset and asked her what was wrong, so Hannah explained what had happened. Then Hannah said "Abbie wasn't feeling well this morning so she went home. Emma can't have been saving the seat for her – she was just making up any old excuse. The real reason that she didn't want me to sit with them is that she's jealous – she thinks her boyfriend Matt has a crush on me, but it isn't true!"

Carolyn told her that she and Emma had been in a meeting all morning, so wouldn't have known that Abbie was off sick. "She probably really was saving the seat for Abbie – she always does. Besides, we sat with them yesterday and that didn't bother her, so I don't think she's jealous of you."

Questions:

Q#		Answer	Type
TOM01	Hannah wanted Carolyn to know that she didn't believe that Matt, Emma's boyfriend, did actually like her.	true	ToM
TOM02	Carolyn thought that Hannah liked Emma's boyfriend Matt.	false	ToM
TOM03	Carolyn told Hannah that Emma had been at training.	true	Memory
TOM04	Emma, who was sitting with Matt, told Hannah that, although she usually saved a seat for Abbie, today she wasn't because Abbie was sick.	false	Memory
TOM05	Carolyn thought that Hannah would realise that Emma had been hoping that Hannah wanted to sit at their table.	false	ToM
TOM06	Hannah, who asked Emma if she could sit with them, sat with Carolyn, because Emma said the seat that was free was for Abbie.	true	Memory
TOM07	Carolyn, who was Hannah's friend, said that Abbie had been in training.	false	Memory
TOM08	Carolyn hoped that Hannah would realise that Emma didn't know that Abbie had gone home.	true	ToM
TOM09	Hannah wanted to sit with Emma.	true	ToM

TOM10	When Hannah came over, Carolyn noticed that she was upset so she asked her what the matter was.	true	Memory
TOM11	Emma thought that her boyfriend Matt realised that Hannah liked him.	false	ToM
TOM12	Emma told Hannah that she couldn't sit with them because the seat which was free was for Abbie.	true	Memory
TOM13	Emma knew that Abbie was sick.	false	ToM
TOM14	Hannah thought that Emma knew that Abbie had gone home sick.	true	ToM
TOM15	Emma told Hannah that Abbie had had to go home sick.	false	Memory
TOM16	Hannah told Emma that Abbie, who was sick, was going to go home, so she didn't need to save her a seat.	false	Memory

Appendix B

Ethics Approval Letter

Human Sciences Subcommittee
Macquarie University, North Ryde
NSW 2109, Australia

22/04/2020

Dear Dr Polito,

Reference No: 52020625515320
Project ID: 6255
Title: Theory of Mind and Emotion Recognition in the Reading in the Mind in the Eyes Test (RMET): a Factor analysis

Thank you for submitting the above application for ethical review. The Human Sciences Subcommittee has considered your application.

I am pleased to advise that ethical approval has been granted for this project to be conducted by Dr Vince Polito, and other personnel: Associate Professor Robyn Langdon, Mrs Wendy Higgins.

This research meets the requirements set out in the National Statement on Ethical Conduct in Human Research 2007, (updated July 2018).

Standard Conditions of Approval:

1. Continuing compliance with the requirements of the National Statement, available from the following website: https://nhmrc.gov.au/about-us/publications/national-statement-ethical-conduct-human-research-2007-updated-2018.

2. This approval is valid for five (5) years, subject to the submission of annual reports. Please submit your reports on the anniversary of the approval for this protocol. You will be sent an automatic reminder email one week from the due date to remind you of your reporting responsibilities.

3. All adverse events, including unforeseen events, which might affect the continued ethical acceptability of the project, must be reported to the subcommittee within 72 hours.

4. All proposed changes to the project and associated documents must be submitted to the subcommittee for review and approval before implementation. Changes can be made via the Human Research Ethics Management System.

The HREC Terms of Reference and Standard Operating Procedures are available from the Research Services website: https://www.mq.edu.au/research/ethics-integrity-and-policies/ethics/human-ethics.

It is the responsibility of the Chief Investigator to retain a copy of all documentation related to this project and to forward a copy of this approval letter to all personnel listed on the project.

Should you have any queries regarding your project, please contact the Faculty Ethics Officer.

The Human Sciences Subcommittee wishes you every success in your research.

Yours sincerely,

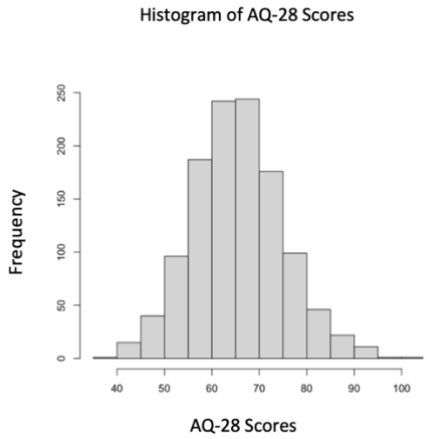
A/Prof Naomi Sweller

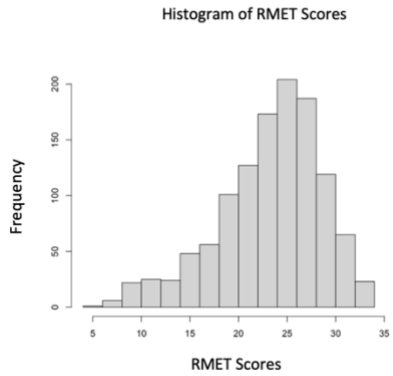
Chair, Human Sciences Subcommittee

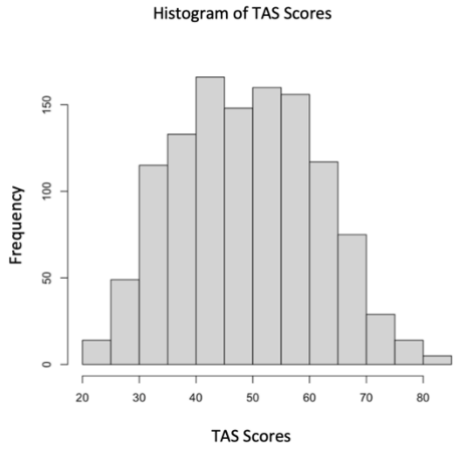
The Faculty Ethics Subcommittees at Macquarie University operate in accordance with the National Statement on Ethical Conduct in Human Research 2007, (updated July 2018), [Section 5.2.22].

Appendix C

Histograms of RMET, TAS, and AQ-28 Scores







Appendix D

RMET Response Frequencies

RMET Question	Target	%	Foil 1	%	Foil 2	%	Foil 3	%
1	playful	71.3	irritated	11.5	comforting	11.0	bored	6.3
2	upset	63.2	terrified	13.5	arrogant	3.9	annoyed	19.4
3	desire	62.8	joking	5.1	flustered	14.2	convinced	17.9
4	insisting	69.1	joking	4.5	amused	12.5	relaxed	13.9
5	worried	78.0	irritated	10.5	sarcastic	9.3	friendly	2.2
6	fantasizing	59.4	aghast	7.8	impatient	**25.8**	alarmed	7.0
7	uneasy	59.1	apologetic	7.0	friendly	23.0	dispirited	11.0
8	despondent	81.6	shy	6.1	relieved	8.8	excited	3.4
9	preoccupied	71.5	annoyed	15.5	horrified	7.7	hostile	5.3
10	cautious	53.1	bored	7.5	insisting	**30.2**	aghast	9.2
11	regretful	65.9	terrified	9.8	amused	16.3	flirtatious	8.0
12	sceptical	72.7	indifferent	16.7	embarrassed	4.9	dispirited	5.6
13	anticipating	65.1	decisive	12.8	threatening	7.9	shy	14.2
14	accusing	57.7	irritated	21.6	depressed	6.3	disappointed	14.4
15	contemplative	62.1	encouraging	10.9	flustered	10.8	amused	14.4
16	thoughtful	65.2	irritated	9.3	encouraging	8.8	sympathetic	16.6
17	doubtful	53.1	playful	14.6	affectionate	**25.8**	aghast	6.5
18	decisive	59.7	aghast	9.4	amused	18.6	bored	12.3
19	tentative	54.3	arrogant	11.0	sarcastic	12.8	grateful	22.0
20	friendly	75.0	dominant	14.4	guilty	8.7	horrified	1.9
21	fantasizing	80.8	embarrassed	6.5	confused	10.0	panicked	2.7
22	preoccupied	66.5	grateful	3.8	insisting	9.6	imploring	20.0
23	defiant	**45.2**	contented	13.4	apologetic	9.0	curious	**32.4**
24	pensive	70.2	excited	4.7	irritated	14.9	hostile	10.1
25	interested	63.3	panicked	6.2	despondent	9.5	incredulous	**29.9**
26	hostile	60.1	alarmed	12.8	shy	9.3	anxious	17.8
27	cautious	68.1	joking	2.4	arrogant	13.3	reassuring	16.2
28	interested	60.1	affectionate	**26.2**	joking	2.9	contented	10.9
29	reflective	66.3	impatient	16.2	irritated	12.8	aghast	4.7
30	flirtatious	85.0	grateful	4.2	hostile	6.0	disappointed	4.8
31	confident	68.2	ashamed	7.0	joking	6.2	dispirited	18.6
32	serious	73.2	bewildered	13.9	ashamed	3.6	alarmed	9.2
33	concerned	65.8	embarrassed	5.4	fantasizing	12.6	guilty	16.2
34	distrustful	56.8	aghast	11.0	baffled	**26.6**	terrified	5.6
35	nervous	**36.3**	puzzled	16.8	insisting	18.4	contemplative	**28.2**
36	suspicious	82.6	ashamed	3.1	nervous	4.3	indecisive	10.1

Note. Values that do not meet the original criteria for inclusion in the RMET (e.g. < 50% select the target, ≥ 25% select the same foil) are indicated in bold.

Appendix E

Tetrachoric Correlation Matrix for the RMET

RMET Item	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
2	0.10																	
3	0.19	0.18																
4	0.08	0.04	0.17															
5	0.12	0.24	0.06	0.18														
6	0.13	0.04	0.21	0.00	0.11													
7	0.00	0.00	-0.04	0.19	0.13	-0.06												
8	0.17	0.23	0.10	0.17	0.21	0.06	0.16											
9	0.04	0.18	0.15	0.09	0.13	0.15	-0.03	0.11										
10	0.00	0.03	0.09	0.03	0.06	-0.01	0.11	0.06	0.09									
11	0.07	0.15	0.13	0.21	0.24	0.01	0.18	0.20	0.11	0.08								
12	0.05	0.18	0.16	0.11	0.12	0.12	0.18	0.17	0.09	0.16	0.17							
13	0.20	0.12	0.13	0.11	0.09	0.18	0.06	0.16	0.16	0.10	0.11	0.23						
14	0.11	0.09	0.05	0.09	0.07	0.09	0.10	0.13	0.12	0.03	0.16	0.23	0.24					
15	0.03	0.15	0.09	0.12	0.16	0.13	0.06	0.24	0.27	-0.03	0.17	0.12	0.14	0.15				
16	0.22	0.13	0.15	0.05	0.20	0.07	0.13	0.13	0.13	0.06	0.11	0.12	0.19	0.21	0.24			
17	-0.08	0.05	-0.03	0.17	0.13	-0.11	0.08	-0.02	0.03	0.08	0.12	0.11	0.10	0.14	0.14	0.16		
18	0.03	0.10	0.07	0.25	0.24	-0.01	0.07	0.20	0.08	0.09	0.27	0.20	0.11	0.20	0.21	0.20	0.07	
19	0.05	0.02	0.14	0.11	0.06	0.04	0.12	0.19	0.17	0.04	0.15	0.14	0.18	0.11	0.17	0.12	0.10	0.07
20	0.29	0.20	0.09	0.00	0.18	0.05	-0.02	0.07	0.17	0.07	0.11	0.11	0.11	0.12	0.15	0.17	0.09	0.07
21	0.22	0.17	0.30	0.12	0.15	0.21	-0.05	0.17	0.28	-0.06	0.11	0.00	0.16	0.17	0.28	0.22	0.01	0.12
22	0.12	0.19	0.10	0.15	0.09	0.06	0.17	0.16	0.31	0.09	0.17	0.18	0.14	0.16	0.18	0.23	0.20	0.12
23	0.09	0.06	0.04	0.15	0.11	-0.03	0.11	0.13	0.14	0.02	0.09	0.13	0.10	0.16	0.17	0.07	0.00	0.10
24	0.12	0.21	0.18	0.03	0.20	0.14	0.09	0.23	0.30	0.10	0.23	0.15	0.19	0.28	0.24	0.21	0.14	0.16
25	0.26	0.09	0.21	0.00	0.08	0.14	0.02	0.11	0.25	0.08	0.07	0.10	0.12	0.03	0.10	0.16	-0.06	0.15
26	0.15	0.15	0.16	0.23	0.10	0.12	0.17	0.25	0.14	-0.03	0.15	0.16	0.14	0.19	0.17	0.06	0.09	0.12
27	0.09	0.09	-0.02	0.14	0.19	0.05	0.16	0.25	0.16	0.09	0.24	0.15	0.15	0.13	0.20	0.20	0.19	0.12

RMET Item	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
28	0.03	0.17	-0.05	0.11	0.12	0.06	0.12	0.09	0.05	0.12	0.09	0.16	0.18	0.21	0.16	0.15	0.16	0.13
29	0.05	0.01	0.09	0.09	0.21	0.20	0.01	0.17	0.22	0.03	0.14	0.18	0.18	0.23	0.12	0.15	0.00	0.19
30	0.29	0.21	0.35	0.23	0.25	0.25	0.14	0.33	0.29	0.08	0.23	0.29	0.26	0.23	0.23	0.24	-0.07	0.21
31	0.09	0.01	0.09	0.11	0.17	0.11	0.05	0.01	0.14	0.02	0.11	0.08	0.05	0.13	0.19	0.12	-0.06	0.23
32	0.09	0.20	0.22	0.18	0.18	0.17	0.02	0.13	0.23	0.04	0.15	0.22	0.21	0.17	0.24	0.24	0.12	0.20
33	0.03	0.06	0.00	0.02	0.12	-0.04	0.03	0.16	0.14	0.13	0.14	0.12	0.12	0.00	0.11	0.11	0.10	0.05
34	0.21	0.19	0.21	0.24	0.15	0.06	0.13	0.21	0.14	0.11	0.25	0.28	0.20	0.19	0.21	0.27	0.10	0.16
35	0.06	0.09	0.06	0.24	0.16	-0.12	0.09	0.11	0.00	0.00	0.13	0.07	0.03	0.07	0.14	0.09	0.15	0.07
36	0.10	0.19	0.13	0.20	0.26	0.06	0.18	0.26	0.18	0.08	0.21	0.23	0.18	0.21	0.15	0.17	0.16	0.09

RMET Item	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36
20	**0.03**																	
21	0.19	0.22																
22	0.09	0.25	0.10															
23	0.06	-0.05	0.09	0.14														
24	0.25	0.20	0.21	0.25	0.09													
25	0.02	0.16	0.21	0.14	0.04	0.11												
26	0.01	0.14	0.26	0.16	0.21	0.09	0.10											
27	0.18	0.03	0.11	0.28	0.15	0.16	0.11	0.11										
28	0.11	0.15	0.03	0.21	0.16	0.11	0.12	0.12	0.13									
29	0.18	0.18	0.15	0.12	0.09	0.29	0.17	0.00	0.13	0.10								
30	0.18	0.23	0.35	0.19	0.15	0.30	0.36	0.31	0.13	0.17	0.34							
31	0.08	0.16	0.36	0.12	0.09	0.18	0.26	0.10	0.06	0.06	0.05	0.19						
32	0.13	0.22	0.30	0.28	0.07	0.25	0.22	0.21	0.16	0.22	0.09	0.19	0.23					
33	0.14	0.09	0.11	0.12	0.07	0.17	0.10	0.01	0.20	0.11	0.02	0.21	0.12	0.21				
34	0.12	0.16	0.10	0.18	0.18	0.23	0.05	0.23	0.15	0.13	0.12	0.21	0.11	0.23	0.12			
35	0.11	0.14	-0.02	0.10	0.09	0.06	0.01	0.16	0.17	0.17	0.00	0.09	0.01	0.10	0.04	0.22		
36	0.09	0.15	0.20	0.18	0.15	0.11	0.23	0.22	0.20	0.23	0.15	0.19	0.16	0.27	0.10	0.24	0.10	

Appendix F

Fit Indices for EFA of the RMET

Number of Factors	CFI	TLI	RMSEA	RMSR	BIC	Chi-square†	Cumulative Variance	Number of Items with Factor Loadings < 0.03
1	0.58	0.55	0.07	0.06	-506	3697	0.14	7
2	0.67	0.63	0.06	0.05	-950	3005	0.18	11
3	0.71	0.65	0.06	0.05	-1021	2693	0.20	9
4	0.71	0.63	0.06	0.05	-873	2607	0.23	10
5	0.77	0.68	0.06	0.04	-1091	2163	0.24	11
6	0.78	0.67	0.06	0.04	-955	2080	0.27	9
7	0.78	0.65	0.06	0.04	-801	2022	0.29	8
8	0.83	0.71	0.05	0.03	-973	1644	0.29	8
9	0.83	0.69	0.06	0.03	-853	1568	0.31	7
10	0.85	0.70	0.06	0.03	-803	1425	0.34	8
11	0.88	0.73	0.05	0.03	-840	1204	0.37	5
12	0.89	0.73	0.05	0.02	-782	1086	0.39	11
13	0.87	0.67	0.06	0.03	-527	1171	0.4	7
14	0.92	0.75	0.05	0.02	-695	840	0.41	9
15	0.92	0.74	0.05	0.02	-588	791	0.41	11

Note. †p-values for chi-square < .001 for all numbers of factors

Appendix G

Comparison of Factor Loadings for High AQ-28 Group with and without Straight Lining Participants

RMET item	Target	Factor 1 Thinking	Factor 2 Negative	Factor 3 Flirting
22	preoccupied	0.652		
32	serious	0.530		
24	pensive	0.511		
9	preoccupied	0.455		
20	friendly	0.352		
28	interested	0.324		
34	distrustful		0.614	
8	despondent		0.568	
5	worried		0.506	
7	uneasy		0.502	-0.372
27	cautious		0.462	
12	sceptical		0.454	
23	defiant		0.449	
4	insisting		0.408	
26	hostile		0.403	
16	thoughtful		0.399	
18	decisive		0.384	
36	suspicious	0.348	0.364	
11	regretful		0.353	
35	nervous		0.344	
21	fantasizing			0.658
30	flirtatious		0.427	0.529
6	fantasizing			0.453
1	playful			0.402
3	desire			0.308

RMET item	Target	Factor 1 Thinking	Factor 2	Factor 3
22	preoccupied	0.583		
28	interested	0.528		
24	pensive	0.374	0.357	
17	doubtful	0.346		
30	flirtatious		0.807	
21	fantasizing		0.766	-0.405
3	desire		0.509	
9	preoccupied		0.502	
32	serious		0.498	
25	interested		0.452	
13	anticipating		0.442	
26	hostile		0.441	
8	despondent		0.440	
29	reflective		0.437	
36	suspicious		0.428	
6	fantasizing		0.405	
1	playful		0.405	
31	confident		0.405	
16	thoughtful		0.400	
20	friendly		0.382	
15	contemplative		0.369	
18	decisive		0.367	
12	sceptical		0.362	
2	upset		0.358	
5	worried		0.329	
7	uneasy	0.329		0.430
34	distrustful		0.369	0.413
4	insisting			0.326

Note. Left table has straight lining participants' data retained