Theory of Mind and Video Games: Developing the Short Story Task-B and Examining Relationships between Theory of Mind and Video Game Play

A Thesis Presented in partial fulfilment of the requirements for the degree of Doctor of Clinical Psychology

Massey University, Palmerston North

New Zealand

Joshua Nathan Robinson

2023

Supervisors:

Dr Aaron Drummond

Dr Michael Philipp

Dr Clifford Van Ommen

ABSTRACT

How video games may affect cognition is poorly understood. This thesis examined one overlooked area – the relationship between video game play and Theory of Mind (ToM), an individual's ability to understand other people's thoughts, feelings, beliefs, desires, intentions, or emotions (Tager-Flusberg & Sullivan, 2000). Complicating the investigation of ToM in this thesis, few measures currently exist that can be used with neurotypical adults, and none have alternate forms.

Across two studies, this thesis aimed to examine whether engagement with different video game genres or social contexts was related to performance on measures of ToM in neurotypical adults. It was also determined whether the General Aggression Model (GAM) or the General Learning Model (GLM) better accounted for observed findings. Further, a new form of the Short Story Task (SST), the Short Story Task-B (SST-B), was piloted. Finally, it was also explored whether literary fiction familiarity was related to ToM ability and whether ToM tests could be substituted for single-item self-report measures.

This thesis found little support for an association between video game play and ToM, a finding which highlighted limitations in both the GAM and GLM. However, there was some support for the notion that literary fiction familiarity was associated with improved ToM. Although the studies developed an incrementally improved ToM measure, like the SST, large issues with the SST-B's psychometric properties were detailed. Results also indicated that single-item self-report measures should not be substituted for measures of ToM. Overall, the limited support for a relationship between video game play and ToM suggests a need for media psychology theories that are broader than just the features of violence. While some support was found for relationships between literary fiction familiarity

and ToM, the psychometric analyses of the SST and SST-B highlight that ToM measures likely require further refinement.

ACKNOWLEDGEMENTS

I always pledged to write the acknowledgements of this thesis last. It feels strange for this time to finally arrive. Thank you to my two primary supervisors, Dr Aaron Drummond and Dr Michael Philipp. Both of you went above and beyond your role as supervisors, offering pastoral support, reassurance, and motivation whenever required. Undoubtedly, without your guidance throughout the Covid-19 pandemic, it would not have been possible to reach this point. 'Thank you' feels insufficient as repayment for all you have offered me. Thank you to Dr Clifford Van Ommen for your presence and encouragement throughout this process and for the participants who offered up their time.

To my parents, Murray and Philippa, your support for me never wavered despite the length of this process. Your frequent check-ins and investment in my general wellbeing did not go unnoticed. Karen and Martin, your genuine interest in my work, rambling emails full of helpful resources, and hospitality have played no small part in making this journey possible. Olivia and Charlene, you are the two people who have been by my side throughout this entire journey. Your humour and comradery have pulled me out of many gullies throughout the DClin. You are great clinicians, and I wish you the best in your careers. To countless others: my brother Ben, aunties, uncles, grandparents, friends (with a special mention to Lauren Hall, whose Zoom calls kept me sane during the pandemic), and clinical staff, you have all given so much of yourselves to make me who I am. Thank you, thank you, thank you.

Last, and certainly not least, a special thank you to my partner, Rebekah. You have been there to lift me up during the hardest times in my life, listening to my long-winded rants and offering nothing but love and practical advice without asking for anything in return. I

look forward to our future together, and I hope I can repay you for the kindness and support you have shown me over these last eight years.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1:

VIDEO GAMES AND COGNITION

Globally, 2.3 billion people play video games (NewZoo, 2019), with recent estimates showing that 67% of New Zealanders are regularly engaging with video games (Brand et al., 2017). Understanding how this may affect cognition is key to allowing individuals to make informed choices regarding media use. Much of the current research has focused on examining how violent video games may influence aggression (Drummond et al., 2020). Therefore, within the present thesis I initially aimed to understand how video games may influence cognition in previously overlooked areas. I then examined existing theories within the video game literature to understand the mechanisms through which video games may cause changes in our cognitive processes. Finally, I explored the current understanding and measurement of one of these cognitive processes, theory of mind.

In this chapter, I employ the Diagnostic and Statistical Manual of Mental Disorders, 5th Edition (DSM-5) framework for neurocognitive functioning to narratively review aspects of cognition that might plausibly be influenced by video game play (American Psychiatric Association, 2013). The primary purpose of this literature review is to a) ascertain what effect video game play has, or might have, upon each cognitive domain and b) identify current gaps in the literature for future research. In addition, this review focused on literature examining a) the effects of commercial games rather than gamified training applications (e.g., brain training) and b) the effects on neurotypical populations. Restriction a) was imposed as the purpose of video game gameplay within brain training studies is fundamentally different (i.e., entertainment versus cognitive improvement and rehabilitation) and warrants an independent review. The effects of video games on the neurocognitive functioning of neurotypical individuals were of interest as this thesis aimed to ascertain areas of clinically relevant

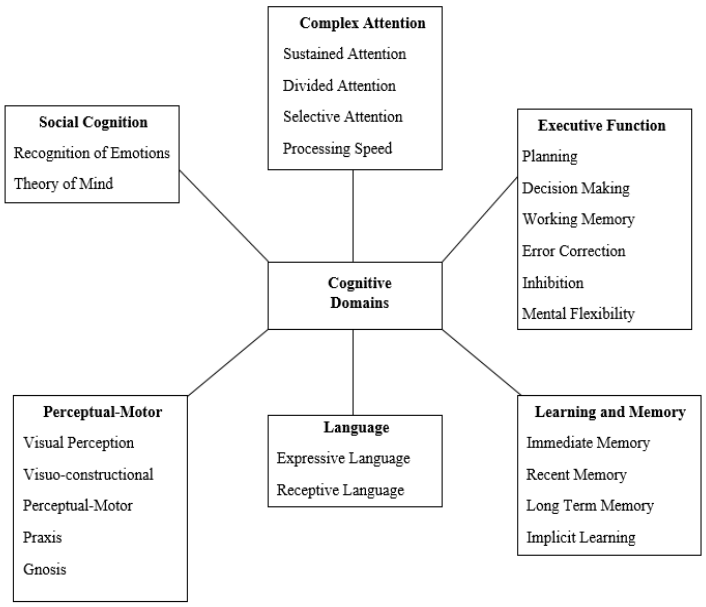
cognitive functioning which might be impaired or improved by the medium, which in turn, may have implications for developing treatments for clinical and subclinical impairments.

Domains of Cognition. The DSM-5 breaks cognitive functioning into six domains. In healthy individuals, these domains function normally; however, deficits in one or more domains can constitute neurocognitive disorders (American Psychiatric Association, 2013). The six domains are complex attention, executive function, learning and memory, language, perceptual-motor, and social cognition (see DSM-5; American Psychiatric Association, 2013, pp. 593-595 for further details). These domains are further divided into subdomains (Figure 1.1). As these domains and subdomains were constructed via expert consensus and show face and field-tested validity, they propose a more robust conceptual model of cognitive function than competing alternatives (Sachdev et al., 2014).

Arguably, the DSM-5 does not provide the best framework for conceptualising cognitive functioning. The DSM-5 has been criticised as unduly categorical and simplistic (Clark et al., 2017). As such, alternative frameworks have been created. Of note is the Research Domain Criteria (RDoC). The RDoC is a dimensional system that breaks cognitive functioning into six inter-related domains with associated subdomains (Morris & Cuthbert, 2012). However, adopting the RDoC for the present thesis is potentially problematic. Notably, a vast majority of the reviewed work was built upon the cognitive model proposed by the DSM-5. Adopting the RDoC in the present thesis would have increased the potential for relevant literature to be omitted. Further, this thesis was completed in partial fulfilment of the requirements for the Doctor of Clinical Psychology qualification. Undoubtedly, the DSM-5 has profound clinical relevance to this qualification and clinical psychology practice. Thereby, the DSM-5 framework has been adopted.

Figure 1.1

The Cognitive Domains Outlined in the DSM-5



Note. Bold text indicates a domain; regular text denotes subdomains.

How Video Games Affect Cognition. Gentile (2011) proposed five elements of video games than can influence cognitive functions: structure, mechanics, amount of play, content, and context. Structure refers to how the information is presented to individuals (e.g., two-dimensional versus three-dimensional). Mechanics refers to how individuals physically interact with these environments using input devices such as keyboards and controllers. Both of these elements are primarily investigated and manipulated in investigating video games' effects on perceptual-motor abilities (e.g., Green & Bavelier, 2006) and likely have limited impact on other cognitive domains.

Alternatively, the impact of play time, content, and context has been investigated in relation to a broader range of cognitive domains. Increased play time is hypothesised to impact cognition through greater repetition of behaviours. However, this also involves

increased interaction with in-game content and contexts. As such, it is more likely that the content and context of video games have the most significant effects on cognition (Gentile, 2011). Content generally refers to what individuals do within the video game (e.g., shoot people). Alternatively, context defines the rules and parameters that govern how the game is played (e.g., trying to capture an in-game objective that may require you to shoot people versus shooting the most number of people possible within a given time). Context can also include additional features, such as whether the game is played by oneself or others (Gentile, 2011). Content and context are commonly included under the term 'genre' (Adams, 2014) and influence cognition by directly controlling the potential learning encounters an individual has when playing a video game. As content and context vary significantly across the wide variety of commercially available video games, contrasting results are often observed within the literature.

In the following section, I review literature examining relationships between video game play and the six cognitive domains outlined in the DSM-5 to identify gaps in the literature that warrant further investigation. To foreshadow, a comprehensive review of the literature reveals that although much is known about several cognitive domains, other domains have either limited or no empirical data about their potential effects. As such, this thesis intends to examine one of these latter domains to understand further the relationships between video game play and cognitive functioning.

Executive Function. Executive function constitutes our ability to conduct a range of higher-order tasks and consists of six subdomains (Table 1.1). Research has explored the effects of video game play across each subdomain. Overall, how video game play may affect executive functioning is inconclusive. Using different video game genres and measures between studies has yielded contrasting results, meaning further research is required to replicate results and clarify effects.

Table 1.1

The Executive Function Domain

Domain/Subdomain	Construct	Operationalisation
Planning	*Predetermination of future actions to complete a goal.* [a]	Ability to find the exit to a maze; interpret a sequential picture or object arrangement.
Decision Making	Deciding in the face of competing alternatives.	Simulated gambling.
Working Memory	Ability to hold information for a brief period and to manipulate it.	Adding up a list of numbers or repeating a series of numbers or words backward.
Error Correction	Ability to benefit from feedback to infer the rules for solving a problem.	
Inhibition	Ability to choose a more complex and effortful solution to be correct.	Looking away from the direction indicated by an arrow; naming the colour of a word's font rather than naming the word.
Mental Flexibility	Ability to shift between two concepts, tasks, or response rules.	[Shifting] from number to letter, from verbal to key-press response, from adding numbers to ordering numbers, from ordering objects by size to ordering by colour.

Note. Constructs in italics are not explicitly defined in the DSM-5. Definitions for these constructs have been interpreted from other sources which use the operationalised measures outlined in the DSM-5. Operationalisation of Error Correction has been intentionally left blank due to this not being defined in the DSM-5 and there being no clear or consistent operationalisation within the literature.

[a] Völter and Call (2014)

Planning. Current findings on whether video game play influences planning abilities are mixed. Correlational findings indicate that increased action video game play is negatively related to planning (Holfeld et al., 2015). Contrastingly, higher self-reported internet gaming has been found to correlate with improved planning abilities (Chen & Hsieh, 2018). Further clouding the issue, experimental research suggests that short-term video game play across various genres (first-person shooter, puzzle, strategy, arcade) does not affect planning ability

(Boot et al., 2008). These findings may suggest potential dosage effects, whereby short-term video game play does not influence planning while long-term use does. Alternatively, differences may be due to comparing different games across different genres and comparing single-player/multi-player play to exclusively multi-player play. Longitudinal studies utilising identical measurement tools may produce more consistent and comparable results, allowing for clearer conclusions about video game play's impact on planning abilities.

Decision Making. Evidence suggests that playing risky racing video games may increase the likelihood of risky decision-making (Fischer et al., 2007, 2009). However, these findings require replication as only two studies have explored this effect. These results may also be exclusive to racing video games, as Buelow et al. (2015) suggested that people assigned to a video game condition made more advantageous long-term decisions on the Iowa Gambling Task than controls. Thus, the content and context of the video game played may differentially influence decision-making abilities.

Working Memory. Research examining the effects of video game play on working memory shows mixed results (Basak et al., 2008; Boot et al., 2008; Powers et al., 2013; Unsworth et al., 2015; Waris et al., 2019). Unsworth et al. (2015) found that self-reported video game experience was unrelated to higher visual or verbal working memory. Similarly, self-reported video game play did not correlate with verbal working memory ability (Boot et al., 2008). In contrast, strategy video game play has been shown to lead to improvements in visual and verbal working memory (Basak et al., 2008). Waris et al. (2019) also observed improvement in visuospatial working memory using Unsworth's procedure after substituting Unsworth's estimate-based video game play measure with a newly validated measure. While a meta-analysis conducted by Powers et al. (2013) concluded that video game play was not associated with overall improvements in working memory, this meta-analysis collapsed visuospatial and verbal working memory tasks. Therefore, contrasting results may be

attributable to video game play improving visuospatial working memory but not verbal working memory.

Error Correction. Literature investigating relationships between video game play and error correction is sparse. Only two studies have investigated the effects of video game play on error correction per se. Schenk et al. (2017) conducted a correlational study investigating the effects of video game play on performance on the weather prediction task (a probability-based exercise where participants must classify one to three cards into one of two groups with the expectation that, through feedback on the accuracy of their groupings, their performance should improve across trials). Using this task, it was observed that video game play might result in improved error correction. Alternatively, Dindar (2018) observed that video game play was not associated with high schoolers' performance on complex problem-solving tasks. Given this paucity of research and contrasting results, much more work is required in this area.

Inhibition. Enough research has been conducted concerning inhibitory abilities to allow for meta-analytic summaries of the data. Meta-analytic data suggest that general video game play is associated with small improvements in inhibitory abilities (Powers et al., 2013). Differential changes in inhibitory abilities may also be observed depending on the video game genre (Engelhardt et al., 2015; Oei & Patterson, 2014; Powers et al., 2013). For instance, Engelhardt et al. (2015) observed that increasing difficulty in a first-person shooter game led to decreased inhibition, while Oei and Patterson (2014) found that increasing difficulty in a puzzle game led to improved inhibition. Contrastingly, Goldstein et al. (1997) described no improvement in inhibition following puzzle video game play. As Goldstein et al. (1997) and Oei and Patterson (2014) employed different games, differences in content and context of the game employed may account for these contrasting results. While (Basak et al.,

2008) observed a positive trend in inhibitory abilities following strategy video game play, this result was non-significant.

Mental Flexibility. A significant body of literature has examined the relationship between video games and mental flexibility. Correlational, experimental, and meta-analytic evidence unanimously suggests that playing video games may improve mental flexibility between predictable tasks (Basak et al., 2008; Boot et al., 2008; Green et al., 2012; Powers et al., 2013). However, action and strategy video game play was not associated with random task-switching improvements (Boot et al., 2008; Oei & Patterson, 2014). In contrast, puzzle video game play was related to improvements in random task switching. Thus, while video game play improves predictable task switching, the genre may moderate the effect of gameplay on random task switching.

Perceptual-Motor. The perceptual-motor domain encompasses our ability to visually perceive stimuli to conduct purposeful motor movements with this information and contains five subdomains (Table 1.2). Research on video game play and perceptual-motor abilities is extensive. Video games appear to improve perceptual-motor abilities, though further research is needed within specific subdomains.

Table 1.2

The Perceptual-Motor Domain

Domain/Subdomain	Construct	Operationalisation
Visual Perception	*Interpretation of information within the visual field in regard to size, orientation, and shape.* *Attending to the visual field.* [a]	Line bisection task; motor-free perceptual tasks [that] require the identification and/or matching of figures; some require the decision of whether a figure can be "real" or not based on dimensionality.
Visuo-Constructional	Assembly of items requiring hand-eye coordination.	Drawing, copying, and block assembly.
Perceptual-Motor	Integrating perception with purposeful movement.	Inserting blocks into a form board without visual cues; rapidly inserting pegs into a slotted board.
Praxis	Integrity of learned movements.	Imitate gestures (wave goodbye) or pantomime use of objects to command.
Gnosis	Perceptual integrity of awareness and recognition.	Awareness of faces and colours.

Note. Constructs in italics are not explicitly defined in the DSM-5. Definitions for these constructs have been interpreted from other sources which use the operationalised measures outlined in the DSM-5.

[a] McIntosh et al. (2017)

Visual Perception. Action video games require players to visually track multiple objects. Accordingly, meta-analyses suggest that action video games may have small-to-medium improvements in visuoperceptual abilities (Powers et al., 2013; Wang et al., 2016). Action video game play is associated with larger central and peripheral visual fields (D. Buckley et al., 2010), improvements in visual attention (Castel et al., 2005; Green & Bavelier, 2003, 2006), greater visual resolution (Green & Bavelier, 2007), and better contrast sensitivity (Li et al., 2009). However, Murphy and Spencer (2009) failed to replicate Green and Bavelier (2003) using a similar task and sample, suggesting further direct and conceptual replication of Green and Bavelier's findings is required. Additionally, current understanding

is primarily limited to correlational findings with action video games. Thus, further experimental research utilising different video game genres would be valuable to further elucidate the relationship between video game play and visual perception abilities.

Visuo-Constructional. In the only extant investigation of video games and visuo-construction abilities, David (2012) observed that playing video games that required visuo-constructional abilities resulted in a non-significant trend toward improved scores on a block design test. Given that the positive trend occurred after 6 hours of gameplay, more intensive long-term interventions may produce significant results. Overall, further evidence is needed to understand how video game play influences visuo-construction abilities.

Perceptual-Motor. Improvements are observed in perceptual-motor abilities as a result of video game play. Meta-analytic results suggest video game play is associated with moderate improvements in perceptual-motor abilities (Powers et al., 2013), improvements in visual-motor tracking (Griffith et al., 1983), and enhancements in laparoscopic and *general* surgical skills (Lynch et al., 2010; Ou et al., 2013). However, Gagnon (1985) reported no relationship between video game performance and hand-eye coordination. Further, video game play is not associated with the improved acquisition of *robotic* surgical performance (Harper et al., 2007; Lynch et al., 2010). This may be due to video games changing three-dimensional motions into two-dimensional inputs. Therefore, virtual reality video games with three-dimensional control may improve *robotic* surgical performance and warrant investigation.

Gnosis and Praxis. No research has investigated video games and gnosis and praxis. As neurotypical adults show limited variation in this subdomain (Ewen et al., 2016), this is not expected to be a fruitful area of research in healthy adult populations. However, there may be value in exploring whether video game play may assist in compensating gnosis and

praxis deficits in neuro-compromised adults. Point-and-click puzzle games (e.g., *The Room*) utilise gnosis abilities by requiring players to find, interact with, and combine objects in an environment to solve puzzles. Using these games in virtual reality plausibly requires the use of praxis abilities. Investigation utilising these games and technologies with these populations may prove worthwhile.

Complex Attention. Complex attention comprises abilities that allow us to maintain and divide attentional resources and contains four subdomains (Table 1.3). It must be noted that this definition differs from that used in diagnosing Attention Deficit Hyperactivity Disorder (ADHD). Much research investigates links between video game play and ADHD symptomatology. However, as ADHD does not employ definitions in the neurocognitive domains, it is beyond the scope of the current review. Generally, the impact of video game play on complex attention is positive. However, as most research has focused on action video games, investigating other genres is warranted.

Table 1.3

The Complex Attention Domain

Domain/Subdomain	Construct	Operationalisation
Sustained Attention	Maintenance of attention over time.	Pressing a button every time a tone is heard, and over a period of time.
Selective Attention	Maintenance of attention despite competing stimuli and/or distractors.	Hearing numbers and letters read and asked to count only letters.
Divided Attention	Attending to two tasks within the same time period.	Rapidly tapping while learning a story being read.
Processing Speed	*The speed at which mental processes can be completed.* [a]	Time to put together a design of blocks; time to match symbols with numbers; speed in responding, such as counting speed or serial 3 speed.

Note. Constructs in italics are not explicitly defined in the DSM-5. Definitions for these constructs have been interpreted from other sources which use the operationalised measures outlined in the DSM-5.

[a] Salthouse (2000)

Sustained Attention. Experimental research indicates that 50 hours of action video game play may improve sustained attention (Dye et al., 2009). Alternatively, correlational evidence shows that action video games do not significantly improve sustained attention compared to other genres (Unsworth et al., 2015). Together, these results suggest that video game play may improve sustained attention irrespective of genre. Contrastingly, Trisolini et al. (2018) found that long-term action video game play was associated with poorer sustained attention over time in adolescents. However, Wolfe et al. (2014) experimentally demonstrated that impaired sustained attention in adolescents due to action video game play was entirely mediated by sleep displacement. Though Wolfe et al. only investigated the short-term impacts of video game play, Trisolini et al.'s failure to control for sleep is problematic. Thus,

while video game play may improve sustained attention, the characteristics of video games implicated in these changes remain unclear.

Selective Attention. Green and Bavelier (2003) provided early evidence suggesting that action video game play was associated with improvements in visual selective attention after only 10 hours of gameplay. Green and Bavelier's findings have subsequently been successfully replicated (Feng et al., 2007). Action and puzzle video game play has also been associated with selective attentional improvements in older adults (Belchior et al., 2013). Further, action video game play is associated with improvements even when using different measures of selective attention (Castel et al., 2005; West et al., 2008). However, Boot et al. (2008) and Murphy and Spencer (2009) failed to replicate these findings. Though contrasting results may be partially due to differences in video game expertise (Boot et al., 2008), Murphy and Spencer recruited a larger sample with similar demographic characteristics to Green and Bavelier. Thereby, direct and conceptual replications of Green and Bavelier's (2003) findings are needed to clarify the relationship between video games and selective attention.

Divided Attention. Greenfield et al. (1994) observed that video game play was associated with improved divided attention. These findings have been replicated using non-action video games after only six hours of gameplay (Satyen, 2005). Some research suggests that action video game players also show a greater ability to perform two tasks concurrently (Strobach et al., 2012; Wu & Spence, 2013). Though Murphy and Spencer (2009) demonstrated no link between video game play and improved divided attention, reaction time was not employed as a primary independent variable (cf., Greenfield et al.), potentially explaining contrasting results. Additionally, three hours of video game training did not improve divided attention in older adults (Seçer & Satyen, 2013), and correlational studies also suggest that video game expertise is not associated with multitasking improvements

(Donohue et al., 2012). While contrasting results may be due to methodological and sample characteristic differences between studies, these conflicting results highlight the need for further research in this area.

Processing Speed. Meta-analyses show that action video game play is associated with improved processing speed without compromised accuracy (Bediou et al., 2018; Dye et al., 2009; Wang et al., 2016). However, these results should be interpreted cautiously, given the small number of extant studies. Additionally, some experimental studies have reported no relationship between action video game play and processing speed (van Ravenzwaaij et al., 2014). Divergent effects may be due to condensed rather than distributed long-term video game play (Bediou et al., 2018). As such, evidence predominantly suggests that action video game play improves processing speed. However, further research should examine other video game genres.

Language. Language is comprised of two subdomains, expressive and receptive language, with expressive language further divided into four underlying abilities (Table 1.4). Currently, only research examining the potential influence of video game play on written receptive language has been conducted. Given this paucity of research, future research avenues are explored.

Table 1.4

The Language Domain

Domain/Subdomain	Construct	Operationalisation
Expressive Language	*Use of verbalisations and/or written language to convey ideas.* [a]	
Naming		Identification of objects or pictures
Word Finding		
Fluency	*Lexical access ability.* [b]	Name as many items as possible in a semantic [e.g., animals] or phonemic [e.g., words starting with "f"] category in 1 minute
Grammar and Syntax	Use of articles, prepositions, auxiliary verbs.	Errors observed during naming and fluency tests are compared with norms to assess frequency of errors and compare with normal slips of the tongue.
Receptive Language	Comprehension [of written and spoken language].	Word definition and object pointing tasks involving animate and inanimate stimuli. Performance of actions/activities according to verbal command.

Note. Constructs in italics are not explicitly defined in the DSM-5. Definitions for these constructs have been interpreted from other sources which use the operationalised measures outlined in the DSM-5. Operationalisations and definitions for some constructs have been intentionally left blank due to these not being defined in the DSM-5 and there being no clear or consistent operationalisation or definition of these within the literature.

[a] Kontiola et al. (1990); [b] Shao et al. (2014)

Written Receptive Language. Drummond and Sauer (2014) showed that video game play was not associated with poorer written receptive language abilities in a large sample of adolescents. In contrast, Hartanto et al. (2018) suggest that weekday but not weekend video game play is associated with poorer reading abilities in adolescents suggesting that a deleterious effect of gameplay on written language may exist. However, more recently, Drummond and Sauer (2019) show paradoxical weekday time-of-day effect associations (i.e.,

reductions associated with before-school but not after-school gameplay) that appear to be more consistent with a third variable explanation. Borgonovi's (2016) research further suggests that video game play may actually be associated with acquiring skills that specifically aid adolescent males' digital reading abilities.

Borgonovi's (2016) findings offer an interesting pathway to future research. Specifically, whether these findings are related to video games generally. A pre-test of reading abilities on both paper and computer, a longitudinal intervention of video game play across various genres, followed by subsequent post-testing, would allow for causal conclusions to be drawn. Additionally, analysis of score differences between males and females could determine whether these benefits are seen exclusively in males. If significant differences are observed, these findings have important implications for assessing adolescents' reading abilities in educational settings.

Language acquisition is most rapid and prolific during early childhood. A child's environment is a crucial factor in the acquisition rate (Larson et al., 2020). The role of video games in this context is currently unclear. It may prove fruitful to investigate whether video game exposure at this time may positively or negatively affect language acquisition and whether such effects are moderated by dose or genre. Plausibly, genres that include a higher degree of spoken language (e.g., adventure) may facilitate more rapid language acquisition. Overall, given the paucity of research within this area, further investigation regarding the potential effects of video game play on language is required.

Learning and Memory. Learning and memory are divisible into four subdomains: immediate memory, recent memory, long-term memory, and implicit learning. Recent memory has three further underlying abilities, and long-term memory has two (Table 1.5). Literature examining the effects of video game play on learning and memory is scarce. While

research indicates that playing violent video games may increase the retrieval of aggressive semantic memories (Barlett et al., 2008), and playing prosocial video games may alternatively increase the retrieval of prosocial semantic memories (Greitemeyer & Osswald, 2011), the broader implications of video game play on memory are unknown.

Table 1.5

The Learning and Memory Domain

Domain/Subdomain	Construct	Operationalisation
Immediate Memory	*Repetition of immediately presented information.* [a]	Repeat a list of words or digits.
Recent Memory	The process of encoding new information.	[Encoding] word lists, a short story, or diagrams.
Free Recall	*Unaided recall of encoded information.* [b]	The person is asked to recall as many words, diagrams, or elements of a story as possible
Cued Recall	*Aided recall of encoded information.* [b]	Examiner aids recall by providing semantic cues such as "List all the food items on the list" or "Name all of the children from the story".
Recognition Memory	*The ability to judge whether presented information has been previously encountered.* [g]	Examiner asks about specific items—e.g., "Was 'apple' on the list?" or "Did you see this diagram or figure?"
Long-Term Memory		
Semantic Memory	Memory for facts.	
Autobiographical Memory	Memory for personal events or people.	
Implicit Learning	Unconscious learning of skills.	

Note. Constructs in italics are not explicitly defined in the DSM-5. Definitions for these constructs have been interpreted from other sources which use the operationalised measures outlined in the DSM-5. Operationalisation of some constructs has been intentionally left blank due to these not being defined in the DSM-5 and there being no clear or consistent operationalisation of these within the literature.

[a] Goh and Pisoni (2003); [b] Malmberg et al. (2014)

Semantic Memory. Drummond and Sauer (2014) showed that video game play was not associated with adolescents' impaired academic performance in science or mathematics. This information is commonly stored in semantic memory (Martin, 2009), implying that video games are unlikely to affect semantic memory. Ferguson's (2015) meta-analytic findings align with Drummond and Sauer's, while other meta-analyses have asserted contrasting results (Adelantado-Renau et al., 2019). Additionally, a greater frequency of video game play has been associated with a poorer understanding of effective strategies to encode information into long-term memory (Drummond & Sauer, 2015). This may result from weekday video game play before school and not video game play in general, suggesting a third variable association (Drummond & Sauer, 2015; Hartanto et al., 2018).

All research to date within this area has been correlational and indicated that video games generally are not associated with decrements in the semantic memory abilities of adolescents. Current findings suggest four pathways to future research: a) the establishment of causal relationships, b) the effects of specific video games or video game genres, c) the influence of video games on other learning and memory abilities, and d) the effects of video game play on other populations.

Regarding a), b), and c), experimental investigations utilising puzzle games may prove worthwhile. Puzzle video games commonly introduce players to a set of rules and then require them to remember and apply these across various settings (Adams, 2014). Plausibly, this requires immediate, recent, or semantic memory dependent upon the time frame between the establishment of the rules and puzzle completion. If these effects emerge, follow-up studies could establish whether puzzle games that require rule remembrance and application (e.g., *The Witness*) are associated with memory improvements compared to puzzle games that do not (e.g., *Tetris*).

Addressing d), investigating video games' influence on memory with older adults may prove valuable. Plausibly, rule-based puzzle games may buffer against age-related memory declines as learning new skills may improve memory in older adults (Park et al., 2014), potentially implicating video games as a pathway to buttress against waning memory performance.

Social Cognition. Social cognition is divisible into two subdomains (Table 1.6). This domain encompasses abilities used to facilitate effective social interactions. Results investigating links between social cognition and video game play are mixed and may reflect the use of different video games and measures between studies. Further investigation of relationships between video game play and this domain is required.

Table 1.6

The Social Cognition Domain

Domain/Subdomain	Construct	Operationalisation
Recognition of Emotions	[An individual's ability to] identify emotion in … faces.	Identification of emotion in images of faces representing a variety of both positive and negative emotions.
Theory of Mind	Ability to consider another person's mental state (thoughts, desires, intentions) or experience.	Story cards with questions to elicit information about the mental state of the individuals portrayed, such as "Where will the girl look for the lost bag?" or "Why is the boy sad?"

Emotion Recognition. Violent video game play has been associated with faster identification of angry faces but slower identification of happy faces, compared to non-violent video game play (Kirsh & Mounts, 2007). Follow-up investigations have failed to replicate these findings (Pichon et al., 2018), only partially replicated findings (Bailey & West, 2013) or have observed alternative findings altogether (Diaz et al., 2016).

The lack of heterogeneity between these studies makes the current understanding of the effects of video game play on emotion recognition unclear. Notably, Kirsh and Mounts (2007) employed the only experimental paradigm within this area. Future research may look to initially replicate these findings given limited evidence for causal relationships within this area. The graphical complexity of video games has also significantly improved over time, so replicating these findings using newer violent video games with increased graphical fidelity (e.g., *Callisto Protocol*) may show differential effects.

Investigation of dosage effects may also prove fruitful. Bailey and West (2013), Diaz et al. (2016), and Pichon et al. (2018) all examined the long-term influence of video game play on emotion recognition, while Kirsh and Mounts (2007) examined effects after only 15 minutes of play. Longitudinal research is required as potential long-term impairment in emotion recognition poses a more severe consequence of video game play.

Theory of Mind. Only two studies have examined video games' effect on theory of mind. The first study investigated whether the inclusion of narration in video games leads to short-term improvements in affective theory of mind (Bormann & Greitemeyer, 2015). Supporting their hypothesis, Bormann and Greitmeyer found that after twenty minutes of video game play, participants who played a video game that included narration showed improvements in affective theory of mind relative to control groups. Alternatively, Kühn et al. (2019) longitudinally investigated whether daily violent video game play for two months led to changes in affective theory of mind. Following the two-month intervention, Kühn et al. did not observe any significant differences in affective theory of mind abilities between the violent video game group and control groups.

These studies provide limited opportunity for generalisation or future research as violence and narration are not present in a wide variety of video games. Video game genres

constitute a higher order of classification, which may guide the future investigation of different effects on theory of mind. Plausibly, strategy games may improve theory of mind as they require players to predict their opponent's moves. Roleplaying games may also be expected to improve theory of mind as the players simulate the mental state of their character, an effect observed in readers of literary fiction (Dodell-Feder & Tamir, 2018). If improvements in theory of mind are causally related to playing a specific genre, then future studies are warranted to investigate the influence of more specific gameplay factors.

Video Games, Cognition, and the Current Thesis. Literature investigating the effects of video game play on the cognitive domains is primarily lacking regarding social cognition, memory, and language. Given the scope of this thesis, these gaps in our understanding cannot all be investigated. As such, this thesis will focus on the effects of video game play on one subdomain of social cognition in neurotypical adults, Theory of Mind (ToM). Specifically, ToM was chosen as it looked to be a particularly fruitful area, given the lack of research. In addition, this thesis is being completed in partial fulfilment of the Doctor of Clinical Psychology qualification, which requires students' research topics to have clinical relevance. ToM has profound clinical relevance, with a range of neurological, developmental, and psychological disorders characterised by supposed deficits in this cognitive ability (Cotter et al., 2018). If it is observed that video game play is related to impairments in ToM, this may provide insight into risk factors for developing ToM deficits in clinical disorders. Alternatively, if improvements are observed, this could offer a starting point for potential treatment interventions or the development of protective mechanisms.

This review initially highlighted that video games' content and context (commonly subsumed under the superordinate term ‘genre’) have the most significant influence on cognition. Thereby, examining the relationship between engagement with different video game genres and ToM abilities offers the most fruitful pathway to further knowledge within

this area. Given the paucity of research within this area, this thesis initially aimed to undertake a correlational investigation of the relationship between ToM and video game genre engagement, as this offered a time and resource-sensitive pathway to determine whether a relationship between these two variables exists. If relationships were observed, further resource-intensive experimental research would be conducted to establish whether causal relationships between these variables are present. In this situation, a pretest-post-test experimental design would be adopted as this offers a stronger establishment of evidence for causality relative to other methods (Salkind, 2010). To foreshadow, a general lack of statistically significant associations in Study 1, difficulties in developing an alternative form of the ToM measure, and the Covid pandemic resulted in the aims of the thesis shifting. Following the completion of Study 1, this thesis instead looked to further investigate the psychometric properties of the employed ToM measures, determine whether tests of ToM ability could be substituted for single-item self-report measures, and examine whether literary fiction familiarity was related to ToM ability.

CHAPTER 2:

THEORETICAL FRAMEWORKS

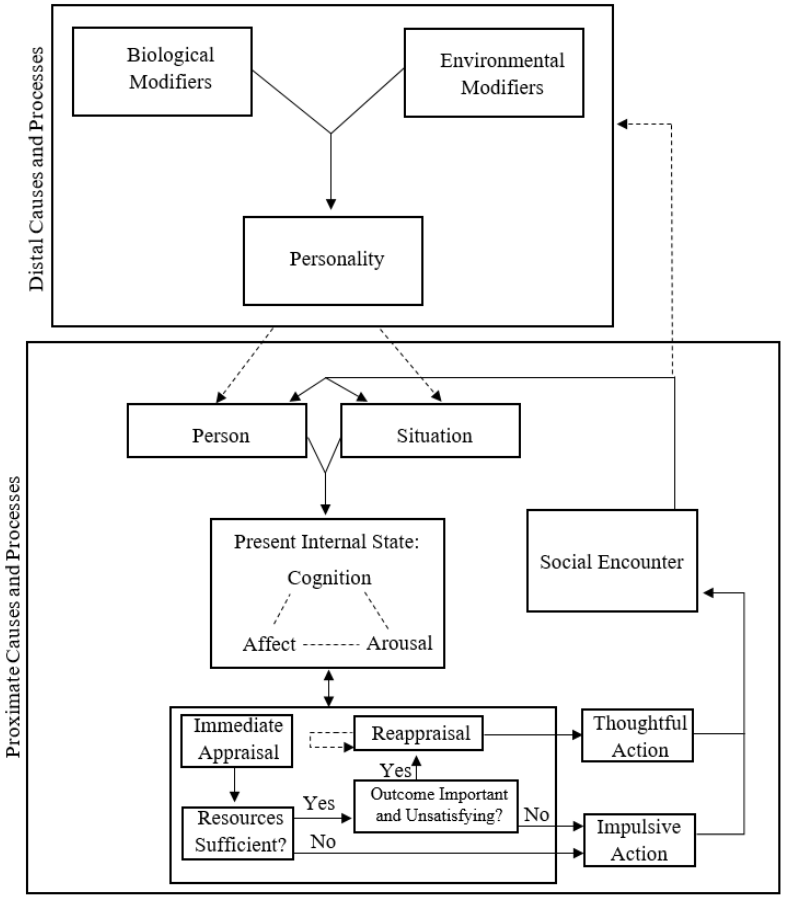
Chapter 1 has detailed that we currently have a limited understanding of the impact that video game play may have on the neurocognitive functioning of neurotypical adults, with a notable dearth of research regarding the potential impact on Theory of Mind (ToM). The paucity of research investigating links between video game play and neurocognitive functioning poses another problem. Namely, existing theoretical frameworks have not been extended to make predictions about the plausible effects of video game play on cognitive functioning outside the influence of violent video game play on aggression. To bridge this gap, this chapter will articulate two of the dominant theories within the video game literature that attempt to explain causal chains in cognition, behaviour, and emotion: The General Aggression Model (GAM) and the General Learning Model (GLM). Although both theories have received criticism in recent years, they remain two of the more commonly adopted theoretical models within the field (Devilly et al., 2017, 2021; Ferguson & Dyck, 2012). These theories will be discussed to a) outline how they propose video game play would potentially be related to changes in cognition, emotion, and behaviour, b) determine their strengths and limitations, and c) to ascertain the predictions these theories make about how video game play may impact ToM to inform their utility in the conduction of present and future research.

The General Aggression Model. The GAM was originally developed by Anderson and Bushman (2002) to explain factors contributing to aggression production in any context. The GAM is now frequently used to predict connections between violent video game play and increased aggression (Scharrer et al., 2018). Given the GAM's focus on predicting changes in aggression, its utility in predicting changes in other cognitive processes and abilities may be somewhat limited.

The GAM acknowledges the role of genetic, developmental, cognitive, and social factors in the causal production of aggression by drawing upon cognitive neoassociation theory, social learning theory, excitation transfer theory, social interaction theory, and script theory (Allen et al., 2018). The mechanisms through which the GAM predicts aggression is induced can be roughly divided into two categories: proximate and distal processes (Allen et al., 2018). Proximate processes are divided into three subprocesses: inputs, routes, and outcomes (Figure 2.1).

Figure 2.1

The General Aggression Model



Note. From "The General Aggression Model," by J. Allen, C. Anderson, and T. Bushman, 2018, *Current Opinion in Psychology, 19*, p. 76. (https://doi.org/10.1016/j.copsyc.2017.03.034). Copyright 2015 by Elsevier Science & Technology Journals. Adapted with permission.

Inputs consider how both person and situational factors contribute to the production of aggression through their impact on cognition, affect, and arousal (Anderson & Bushman, 2002). Person factors are individual characteristics that increase or decrease an individual's likelihood to aggress (e.g., trait aggression). Situational factors are features of a situation that may provide cues to aggress, such as frustration due to thwarting competence. Both situational and personal factors work independently to influence an individual's cognition, affect, and arousal.

The second stage of the GAM, routes, proposes how these aforementioned inputs change the individual's cognition, affect, and arousal, referred to as an individual's "present internal state" (Anderson & Bushman, 2002, p. 38). These three variables are interrelated, meaning changes in one may induce changes in the others (Allen et al., 2018). For instance, the GAM postulates that certain situations, such as violence in a video game, may prime an individual to think aggressively. The GAM also predicts that continuous exposure to these situations may lead to the formation of aggressive scripts – cognitions and schemas – resulting in aggressive cognitions becoming more readily accessible over time. The model also postulates that this likelihood may be increased by person factors (e.g., temperament, trait aggression). Similarly, under the model, input variables can readily influence our emotions and moods, potentially resulting in aggressive affect. Finally, situational and person factors are postulated as having the ability to increase or decrease arousal. Unrelated arousal may be misinterpreted for anger and strengthen one's likelihood to aggress (Zillmann, 1988). Alternatively, increases in arousal may plausibly increase aggression (Allen et al., 2018).

According to the GAM, one's present internal state is thought to directly influence how the individual reacts to a given situation or social encounter. If an individual has an aggressive internal state, they may be more inclined to immediately appraise a situation as warranting aggression (Anderson & Bushman, 2002). If the individual does not have time to

reappraise the situation, then action will be taken as a direct result of this appraisal. Alternatively, reappraisal may occur, resulting in a change in one's internal state and the enactment of a non-aggressive behaviour (Allen et al., 2018). Irrespective of the course of action taken and the resulting outcome, following action, the individual receives social feedback, which may influence the input factors in positive or negative ways dependent upon the action taken. Given this cyclical nature, repeated violent video game play may, according to the GAM, lead to long-term changes in one's knowledge structures, increasing an individual's likelihood to aggress (Barlett et al., 2009). Distal processes are enduring variables hypothesised to alter input factors (Anderson & Carnagey, 2004). These factors, such as an individual's cultural background or genetic makeup, influence the situations a person encounters and thereby shape their personality (Allen et al., 2018).

Criticisms of the General Aggression Model. While the GAM has provided the dominant framework for the investigation of cognitive, emotional, and behavioural changes in the context of violent video game play, it has also been frequently criticised for a) what some researchers suggest are incorrect assumptions it makes, particularly about aggression, b) its apparent lack of falsifiability, and c) its perceived failure to adequately draw upon the role of distal processes, and the theories that supposedly underpin the GAM, in explaining aggression production (Devilly et al., 2017; Ferguson & Dyck, 2012). Regarding a), Ferguson and Dyck (2012) highlight that the GAM assumes that aggression, while once evolutionarily adaptive, is now universally bad. Alternatively, it is argued that aggression often exists on a continuum of adaptive to maladaptive (Ferguson & Beaver, 2009). Given this assumption that aggression is always maladaptive, the relatively small increases in aggression observed in video game studies (e.g., $r < .20$) are often generalised to account for the enactment of real-world violence. This view potentially discounts the magnitude of the effect size and the

possibility that a small increase in aggression may not lead to maladaptive or extreme consequences, such as violent behaviour.

The GAM also appears to assume that humans cannot differentiate between media and reality (Ferguson & Dyck, 2012). That is, the GAM posits that exposure to violence in media is equivalent to witnessing violence in real life and will thereby affect all individuals negatively. This claim is often supported by drawing links between witnessing media violence and changes in neural activation as evidence of real-world behavioural changes. However, these assertions are often made without causal evidence linking violent media consumption to the enactment of real-world violence (Ferguson & Dyck, 2012). In addition, the claim that humans cannot differentiate media and reality is also disputed, with research showing that children appear able to distinguish between reality and fictional media by as young as age five (Woolley & Van Reet, 2006).

Further, Ferguson and Dyck (2012) contend with the GAM's assumptions that an individual's experience of aggression is primarily learned, experienced cognitively, and occurs automatically. While environmental learning plays a role in aggression production, biological and genetic influences are also strongly implicated. In addition, some literature suggests that general environmental stress, as opposed to witnessing violence, is a stronger predictor of an individual's likelihood to aggress (Barash & Lipton, 2011). Thus, while the GAM is correct in claiming that aggression may be learned, the role of learning processes in aggression production may be less significant than the GAM asserts.

The role of cognitions in aggression production within the GAM primarily rests on research showing that exposure to violent media primes aggressive cognitions. Repeated exposure is posited to result in the formation of aggressive scripts, which supposedly increases an individual's likelihood to aggress in the future. However, priming effects do not

always translate to intent or evidence of script formation (Ferguson & Dyck, 2012). In addition, the assertion that aggression is a primarily cognitive experience does not appear to fit the existence of both reactive and instrumental aggression. While it may be argued that GAM is designed to account for experiences of instrumental aggression, this appears incompatible with GAM's assumption that aggression is a primarily automatic process. Conversely, this assumption of automaticity better fits within accounts of reactive aggression. However, some literature suggests that diathesis-stress models of aggression production are more accurate than the GAM in explaining reactive aggression production (Ferguson et al., 2008). Thus, it could be argued that the GAM's assertions of aggression as primarily automatic and cognitive appear incompatible and thereby require revision.

Regarding the second asserted limitation, the GAM rests on the premise that falsifiability is central to scientific progress and theory revision (Popper, 1983). Once evidence is presented that a theory cannot account for, the theory must be revised. In the case of the GAM, findings that stand in contrast to the theories assertions are often justified as simply reflecting Type II error, allowing the GAM to persist (Ferguson & Dyck, 2012). Publication bias further complicates the issue, whereby research producing results contrasting assumptions made by the GAM are less likely to be published than those in support (Ferguson & Kilburn, 2009). While these issues are often understood and acknowledged within the literature citing the GAM, an absence of competing alternative theories has likely contributed to the GAM's survival.

Alternatively, the GAM may hold utility in its use as a metatheory (Finkel, 2014). Metatheory differ from traditional theory in that their underlying assumptions are held to be true and are not posited to be falsifiable. They thereby intend to serve as a foundation for subsequent theory or research questions to be built upon. However, the GAM is still frequently adopted as a traditional theory (e.g., Quan et al., (2021); Burnay et al., (2022)).

While this does not discredit the utility of the GAM as a plausible metatheory, its ongoing use as a traditional theory further highlights the absence of alternative theoretical frameworks with the media psychology literature. Thus, should the GAM not hold utility in investigating links between video game play and wider facets of neurocognition, it may still plausibly serve at the foundation for subsequent theory to be built upon.

Finally, as outlined in the final asserted limitation, the GAM has been criticized for supposedly failing to draw upon the theories underpinning it. Notably, the GAM claims to draw upon cognitive neoassociation theory, social learning theory, excitation transfer theory, social interaction theory, and script theory (Allen et al., 2018). However, some authors contend that the GAM primarily relies on script theory in explaining aggression production (i.e., scripts stored in memory are activated by environmental stimuli leading to aggression production) while poorly articulating how other theories inform the GAM (Ferguson & Dyck, 2012). Proponents of the GAM refute this, drawing upon the GAM's acknowledgement of personality and biological factors (i.e., distal causes and processes) being related to aggression production. However, explicit links between distal and proximate processes are criticised as often being poorly articulated and seldom elaborated upon within the literature (Ferguson & Dyck, 2012). In addition, other authors criticise the GAM for appearing to construe some personality traits as little more than a series of scripts and schemas, a perspective which they claim is inconsistent with conventionally held understandings of human personality (Gilbert & Daffern, 2011). Thus, despite the GAM being the dominant theory adopted in the video game literature, it appears to pose several shortcomings to adoption.

The General Learning Model. The GLM, developed by Buckley and Anderson (2006), is an extension of the GAM. Where the GAM is primarily utilised to explain how violent content in media may lead to changes in an individual's likelihood to aggress/act

violently, the GLM seeks to propose a model by which any media can result in the learning of any related behaviour, cognition, or emotional response. For example, the GLM would predict that playing a video game that encourages players to act altruistically (e.g., *Death Stranding*) would theoretically result in the formation of related beliefs/scripts, thereby increasing an individual's likelihood to act altruistically in other contexts.

Given that the GAM forms the basis of the GLM, mechanistic pathways to various outcomes outlined earlier are also utilised within the GLM to explain the process of learning through video game play (Figure 2.1, Page 24). However, conceptual differences within these processes allow the GLM to provide predictions for how video games may result in learning unrelated to aggression. Regarding input variables, the GLM emphasises situational and person factors that influence one's capacity to learn (e.g., age, intelligence, self-esteem; Buckley & Anderson, 2006). Additional emphasis is also placed upon aspects of the game itself (Buckley & Anderson, 2006). Where the GAM focuses almost exclusively on the presence of violence, the GLM considers the role of factors such as time of exposure, content, and context and how these may influence learning (Barlett et al., 2009; Buckley & Anderson, 2006). Additionally, the GLM asserts that video games, relative to other forms of media, present situations which are more likely to result in enduring changes by allowing the player to control the level of difficulty; repetition of specific actions/skills; active engagement; immediate feedback based upon the decisions the player makes; and skill generalization across games (Sarmet & Pilati, 2016).

Regardless of whether an impulsive or thoughtful action is taken as a partial result of one's internal state, an individual has a positive or negative learning encounter (as opposed to a social encounter as seen in the GAM). Plausibly, this difference may reflect the GAM's primary focus on aggression production within social contexts (hence, social encounter), while the GLM focuses more broadly on any learned behaviour (hence, learning encounter).

However, this distinction appears largely semantic, with Buckley and Anderson (2006) not elaborating on this difference between models. Highlighting this, Anderson has co-authored publications that appear to use these terms interchangeably (Gentile et al., 2009).

Dependent upon the type of game, this may allow an individual to learn information or behaviours resulting in short-term cognitive or behavioural changes. Over time and through repeated observational learning encounters, knowledge structures are thought to be learned, reinforced, and become more readily accessible in a variety of situations. Biological factors partially mediate the creation of these knowledge structures. Once created, these can change one's perception, affect, behaviour, beliefs, and expectations, resulting in long-term personality changes as a direct result of learning encounters through video game play (Buckley & Anderson, 2006).

Criticisms of the General Learning Model. As the GLM operates on similar mechanistic pathways to those in the GAM, criticisms of the GAM may also be relevant to the GLM. Notably, the GLM has also been criticised for supposedly lacking true criteria for falsifiability and for failing to adequately draw on the theories underpinning it (i.e., it is also purported to be primarily a social script theory; Ferguson and Dyck, 2012). In addition, it has also been criticised for what some researchers see as a failure to adequately elaborate on how distal processes (e.g., biology, genetics, personality) influence learning. Sarmet and Pilati (2016) highlight that the outcomes that occur in the GLM resulting from an individual's present internal state (i.e., whether an individual chooses an impulsive or thoughtful action) also appear to have inadequate evidence to justify their inclusion within the GLM. Additionally, how game content and context supposedly mediate whether an impulsive or thoughtful action is taken remains presently unclear. Some authors argue that insufficient evidence for this aspect of the model brings into question the validity of the assumptions

underpinning the GLM, given that this implicit process supposedly mediates all learning encounters.

The GAM, GLM, Theory of Mind, and the Current Thesis. Despite both models being criticised for what some see as significant limitations, theory generation and related investigation are beyond the scope of a Doctorate of Clinical Psychology thesis. Further, while both theories are widely employed and well-understood in the present literature, their utility for investigating the relationship between video game play and ToM is currently unclear. Here, I will thereby look to test both theories within this thesis. How the GAM and GLM may predict video game play will influence ToM will therefore be outlined to determine which model (if either) offers more accurate predictions and which model may hold greater utility for guiding hypothesis generation in the conduction of future research within this area.

How the GAM predicts video game play may influence ToM has not been clearly articulated in the extant literature. However, the GAM asserts that violent video game players develop an increased likelihood to aggress over time, partially due to reductions in trait empathy (Anderson & Bushman, 2018). Empathy and ToM often work in tandem, employing overlapping brain regions and involving similar cognitive processes (Cerniglia et al., 2019). Certain facets of ToM and empathy (e.g., cognitive empathy and affective ToM) share such conceptual overlap that the two terms are used interchangeably throughout the literature (Rogers et al., 2007). Given that these two cognitive systems commonly require each other to adequately function and share significant conceptual overlap, decreases in empathy would plausibly result in reductions in ToM. Thus, if the GAM were correct, the model would seem to predict that violent video game play would result in decreases in ToM via a reduction in trait empathy.

Alternatively, the GLM allows predictions about changes in any cognitive system based on video game play. Whether these changes are positive or negative depends upon the content and context of the video game (Barlett et al., 2009; Buckley & Anderson, 2006). Plausibly, the GLM would predict that repeated learning encounters in games that reward the player for predicting the mental state of their opponent (e.g., strategy games) would result in increases in ToM. In contrast, games that do not require mental state prediction, but include violent content, would theoretically result in ToM reductions via similar mechanistic pathways to those elaborated for the GAM. However, the GLM's consideration of a video game's overall content and context means that video games that include violence but reward prediction of other humans' mental states (e.g., *Counter Strike: Global Offensive* a first person shooter game) may still plausibly result in a net improvement to ToM. Additionally, the video game's context would also be considered in predicting ToM changes under the GLM. As ToM is inherently a socio-cognitive ability, an individual who played *Counter Strike: Global Offensive* in a multi-player context would be expected to have a more significant net improvement in ToM, relative to an individual playing in single-player modalities, due to repeated learning encounters through discourse and interaction that occur within multi-player video games.

The GAM would contrastingly predict that ToM abilities would decrease in this situation, regardless of the social context. As such, it is currently unclear which theory (if either) offers a more accurate theoretical framework for investigating video games' effects on ToM. Therefore, this thesis will examine whether correlations between video game genre engagement and ToM abilities fit better within accounts made by the GAM or GLM.

CHAPTER 3:

THEORY OF MIND

So far, this thesis has reviewed literature investigating the relationship between video game play and cognition and outlined existing theories within the video game literature. To further understand the relationship between video game play and Theory of Mind (ToM), this chapter will examine ToM as a construct and how it is measured. ToM was initially conceptualised as one's ability to "impute mental states to [oneself] and to others" (Premack & Woodruff, 1978, p. 515). Mental states may be an individual's thoughts, feelings, beliefs, desires, intentions, or emotions (Tager-Flusberg & Sullivan, 2000). Knowledge of what others may be thinking, and the realization that this may differ from our thoughts, is critical in navigating conversations and facilitating social cooperation (Gweon & Saxe, 2013). ToM abilities also underpin our understanding of more complex conversational abilities, such as detecting and understanding sarcasm and irony (Hughes & Leekam, 2004). Thus, ToM abilities are crucial for social cohesion and everyday social interaction.

A range of theories and models positing how ToM's underlying abilities may function and be differentiated have been articulated (Schaafsma et al., 2015). Theoretical models commonly assert that ToM is underpinned by various related but dissociable subprocesses (Schurz & Perner, 2015). These theories affect ToM's measurement, measure selection, and interpretation. As such, this chapter will initially provide an overview of the primary theories and models of ToM. This overview discusses how our ToM system may function, the subprocesses that may underpin ToM, and how these are measured. Once a model of ToM with the greatest empirical support has been identified, measures of ToM that are consistent with this model, and are commonly employed with neurotypical adults, will be examined. This examination will consider these measures' strengths and weaknesses to determine how ToM will be measured in this thesis.

Theory-Theory and Simulation Theory. Theory-theory (Gopnik & Wellman, 2012) and simulation theory (Gallese & Goldman, 1998) have provided the dominant framework for investigating *how* our ToM system may function across the last 40 years (Schaafsma et al., 2015). Both theories seek to explain the cognitive processes which underpin how we engage in ToM (Apperly, 2008). Theory-theory asserts that humans learn an array of concepts (i.e., beliefs, desires) through personal and observational experience. Much like scientific processes of theory revision, these concepts are constantly changing. Repeated observations and experiences consistent with existing concepts lead to these being reinforced. Alternatively, observations or experiences inconsistent with one's existing concepts lead to them being revised. In addition to the development of concepts, people also develop rules about these concepts (e.g., people commonly behave in ways consistent with their beliefs; people will act in ways to meet their desires which are also consistent with their beliefs). Rules and concepts, which may or may not be explicitly accessible, form a 'theory' about how humans' mental states lead to engagement in various behaviours. Then, using this personal 'theory', humans predict the mental states and behaviour of others. In sum, theory-theory asserts that humans learn to predict the mental states of others based on rules and concepts reinforced through observation and experience (Gopnik & Wellman, 2012).

In contrast, simulation theory rests on the premise that all individuals have a biological basis underpinning the formation of mental states. Given that these underlying cognitive processes should occur similarly across all individuals, simulation theory asserts that we engage in ToM by using our minds to simulate the mental states of others (Gallese & Goldman, 1998). That is, we temporarily adopt the beliefs and desires we perceive others to have and predict their behaviour based on these simulated beliefs and desires.

Originally, simulation theory was developed due to a lack of empirical support for theory-theory (Gordon, 1986). However, current evidence supports the presence of both ToM

systems. Neuroimaging implicates the cortical midline structures and medial prefrontal cortex as being involved in the simulation of others' mental states (Mahy et al., 2014). Although this is an issue under debate, mirror neuron systems may also be implicated. Notably, simulation theorists assert that mirror neurons are central in the simulation of other minds. However, evidence suggests that individuals with damage to mirror neuron systems can still engage in mentalisation (Spaulding, 2012). These conflicting findings indicate that while mirror neurons may be implicated in facilitating ToM, their role does not fit with a simulation theorist's account.

Regarding theory-theory, neuroimaging studies primarily implicate the temporoparietal junction (Mahy et al., 2014). However, there is otherwise a scarcity of research regarding the theory-theory neural basis for ToM. Despite the dearth of research in this area, current literature primarily supports a hybrid theory-theory and simulation theory model, whereby both systems may act independently or in tandem (Apperly, 2008; Kühberger & Luger-Bazinger, 2016). However, some authors reject all three of these theories. Notably, these authors draw on a lack of clear distinction between the theories, the under-appreciation of these theories for the cognitive resources ToM likely employs, and simulation theory's lack of appreciation for individual differences (Korkmaz, 2011). Regardless, theory-theory and simulation theory remain widely accepted within the current literature.

Literature adopting theory-theory, simulation theory, or a hybrid theory to investigate ToM abilities commonly utilises neuroimaging or behavioural measurement tools. Deviating from this, Bivona et al. (2018) attempted a psychometric evaluation of theory-theory and simulation theory ToM abilities. Bivona et al. proposed that theory-theory ToM relies heavily on executive functions while simulation theory ToM relies on emotional/affective systems. Thereby, Bivona et al. posited that following a traumatic brain injury, poor performance on

ToM tasks would correlate with poor performance on executive functioning tasks if the theory-theory model of ToM were valid. Alternatively, a relationship between poor performance on ToM tasks and poor performance on various measures of one's emotional experiences (e.g., empathy) would support a simulation theory model of ToM. To investigate this, Bivona et al. collated 32 studies investigating the impacts of traumatic brain injury on these three areas (ToM, executive functioning, emotional functioning) and examined correlations between them. Findings from this study were ambiguous, as relationships between ToM, executive functioning, and emotional functioning emerged within some studies but not others. Ultimately, no firm conclusions about the validity of theory-theory or simulation theory could be drawn. Regardless, correlating measures of ToM, executive functioning, and emotional functioning is not common practice to explore one's theory-theory or simulation theory abilities and highlights the lack of psychometric measures available to assess individual differences in these ToM abilities.

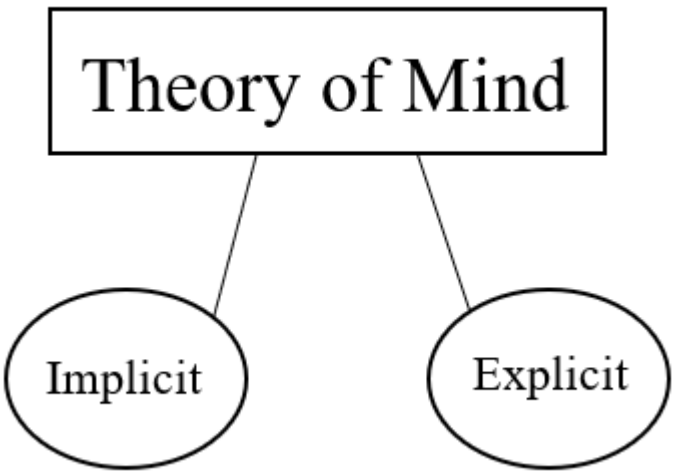
Theory-theory and simulation theory accounts of ToM are primarily investigated using neuroimaging and lack appropriate psychometrics for their measurement (Schurz & Perner, 2015). In the present study, clear operationalisation of ToM is required for empirical investigation of individual differences in ToM abilities between neurotypical adults. Adopting theory-theory, simulation theory, a hybrid theory, or a similar neurocognitive ToM theory is thereby inappropriate. Thus, while theory-theory and simulation theory appear to be valid and supported theories of ToM functioning, an alternative and more readily operationalizable theory is required for the present research.

Implicit and Explicit ToM. While theory-theory and simulation theory offer competing perspectives on how our ToM system may function, implicit and explicit theories of ToM describe non-competing cognitive subprocesses of ToM (Figure 3.1). These cognitive subprocesses are posited to function in tandem with other higher-order ToM

processes, including those proposed by theory-theory and simulation theory (Schaafsma et al., 2015). Generally, implicit ToM is a form of unconscious, unintentional, and automatic mental state attribution that develops during infancy. These attributions are not culturally inherited and are thereby relatively simplistic (e.g., understanding that oneself can have different mental states than others; Kulke et al., 2018; Schneider et al., 2017). Alternatively, explicit ToM is a conscious and purposeful form of mental state attribution that develops during early childhood and advances across adolescence and early adulthood. Given that this process requires conscious awareness, individuals can verbalise their rationale for attributing a particular mental state to another person. Explicit ToM also includes understanding complex social processes such as sarcasm, irony, and faux pas. Resultingly, explicit ToM is likely culturally bound and requires consideration of complex social information in attributions (Schaafsma et al., 2015; Schuwerk et al., 2015).

Figure 3.1

The Implicit-Explicit Model of ToM



Initially, implicit ToM was conceptualised on the basis that children under the age of four were thought not to possess a ToM but readily engaged in behaviours that required some form of mental state attribution (Southgate et al., 2007). Commonly employed ToM measures at the time were relatively cognitively complex, requiring children to attribute a mental state while simultaneously engaging in response selection and inhibition. It was thought that these task demands were too great for most children's underdeveloped temporal and frontal lobes, thereby masking their true ToM ability (Poulin-Dubois & Yott, 2018). As such, new measures were developed to investigate the implicit ToM abilities of children.

The evidence supporting an implicit ToM system has primarily utilised simple behavioural measures with infants and young children. Commonly used are anticipatory-looking false belief measures. These measures usually present children with images showing two actors interacting with an object over a series of temporally ordered images. For instance, initially, Actor One places an object in a box and then leaves the room. This object is moved to a second box by Actor Two without Actor One witnessing this. Actor One then re-enters the room. It is posited that if individuals possess implicit ToM, their gaze should automatically shift to the location where Actor One last saw the object (and therefore would logically believe the object to be), rather than where the object actually is (Kulke et al., 2019). Using this paradigm, Southgate et al. (2007) demonstrated that the majority of 2-year-olds correctly anticipated where Actor One would look for the object after it had been moved. Conceptual replications of this finding using similar experimental paradigms have indicated that implicit ToM abilities may emerge in children as young as 15 months (Onishi & Baillargeon, 2005).

Explicit ToM is measured using any ToM measure requiring conscious judgements about another's mental state. For example, the Faux Pas Test is a widely used ToM task that has individuals view a series of fictional vignettes. These vignettes either display a normal

social interaction or show a scenario in which a character says something to another character that they should not have, without appearing to be aware of their mistake. Identification that a faux pas has occurred requires consideration of both characters' mental states, a process that requires conscious thought (Megías-Robles et al., 2020). While other ToM measures present individuals with different stimuli and response styles, they also primarily require conscious consideration of mental states. Thereby, the majority of commonly utilised ToM measures can be considered explicit measures of ToM. As conscious consideration of mental states is central to measuring ToM throughout the literature, there is little contention regarding the existence of explicit ToM.

Alternatively, there is ongoing debate regarding the existence of an implicit ToM system. Notably, commonly used measures lack adequate validity (Dörrenberg et al., 2018). Resultingly, replications utilising these measures (e.g., anticipatory looking false belief measures) often produce null results, whereby individuals' tendency to look at either box does not occur at levels above chance (Kulke et al., 2019; Kulke et al., 2018). These null results are found across both infant and adult populations. Meta-analytic findings also suggest caution in interpreting research supporting implicit ToM accounts. Barone et al. (2019) found that children consistently selected the correct answer at an above-chance rate on some implicit ToM measures (i.e., violation of expectation tasks) but not others (e.g., anticipatory looking false belief measures). Barone et al. suggest that if children develop ToM before age four, consistent findings should be present across all measures.

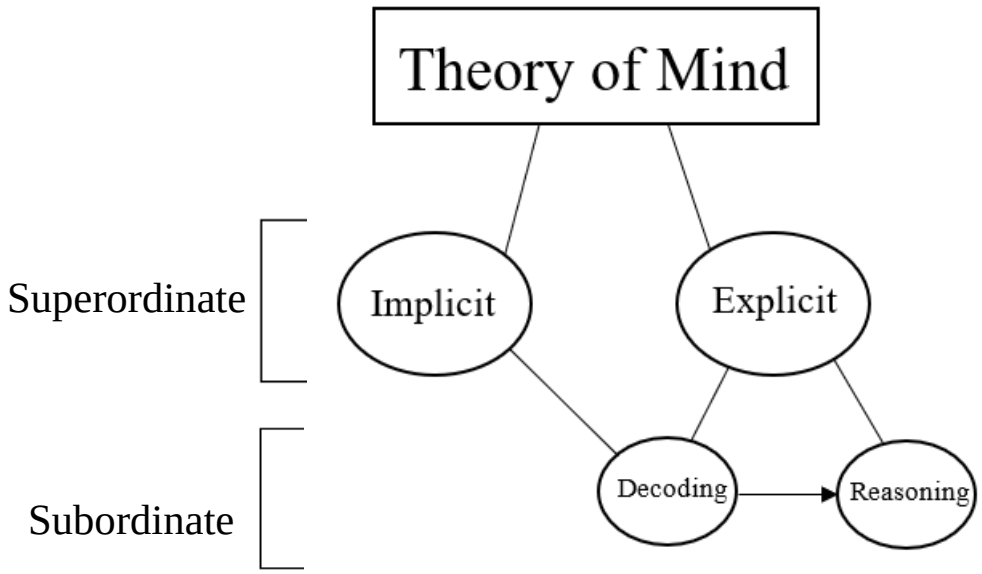
Additionally, implicit ToM studies appear strongly subject to publication bias, whereby studies with small effect sizes often appear to go unpublished (Barone et al., 2019). As such, it is currently unclear whether implicit ToM can be reliably measured, and there is scholarly debate about whether it even exists (Barone et al., 2019). Therefore, it is more prudent within the current thesis to explore and employ measures of explicit ToM in the

investigation of individual differences in ToM abilities between neurotypical adults, given its wider acceptance within the literature, in tandem with practical limitations in the measurement of implicit ToM.

Decoding and Reasoning ToM. Decoding and reasoning ToM abilities are non-competing subprocesses that may underpin our explicit and implicit ToM abilities (Figure 3.2). Decoding ToM is our ability to infer mental states based on immediately present social information (e.g., eyes, tone of voice, facial expressions). This process is generally unconscious, rapid, and automatic, but can be brought to conscious awareness if required (Duclos et al., 2018). Reasoning ToM is a ToM ability which involves conscious cognitive processes. It involves us extrapolating our assessment of the internal mental state of another person to predict that person's subsequent actions or mental state (Dodell-Feder et al., 2013; Sabbagh, 2004). This process may involve accessing previous knowledge about the individual(s) whose mental states are being predicted and integrating this with information about the social context. These two processes often work together in tandem, with ToM decoding potentially facilitating engagement in ToM reasoning (Sabbagh, 2004).

Figure 3.2

The Decoding-Reasoning Model of ToM in Relation to the Implicit-Explicit Model

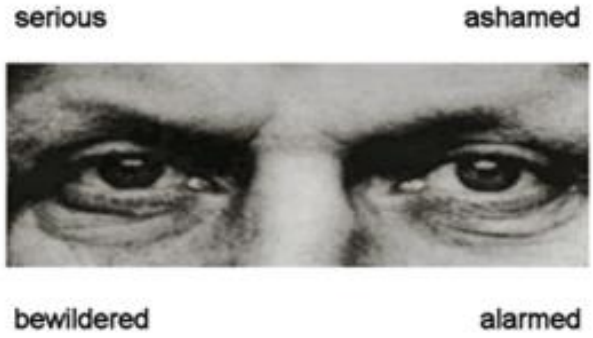


Evidence for dissociable decoding and reasoning ToM abilities has primarily utilised behavioural measures with neurologically or psychologically impaired individuals. Njomboro et al. (2008) provided initial support for this dissociation; it was observed that individuals with left parietal and superior temporal damage (neurological areas thought important for ToM) had impaired performance on reasoning but not decoding ToM tasks. Further, damage to the ventromedial prefrontal cortex (similarly, also neurological areas thought important for ToM) has been associated with specific impairment in ToM reasoning (Geraci et al., 2010). Further supporting this dissociation of the neurological areas implicated, schizophrenia has been associated with significant impairment in decoding ToM but not reasoning ToM (McGlade et al., 2008). It is hypothesised that these differences result from ToM reasoning recruiting diffuse neural networks, while ToM decoding is associated with the specific activation of orbitofrontal and medial temporal circuits (Sabbagh, 2004). The observation that these abilities may be differentially impaired or preserved supports the notion that ToM decoding and reasoning exist and may act as independent systems.

While the majority of currently employed ToM measures were not originally created to measure decoding and reasoning ToM, they commonly employ tasks requiring both of these ToM abilities. Exemplifying this, the Reading the Mind in the Eyes test (Figure 3.3) was initially developed to discriminate between healthy adults and individuals with autism spectrum disorder on *general* ToM ability (Baron-Cohen et al., 2001). This test has individuals view images of people's eyes and then select one of four mental state descriptors that best represent the emotion displayed. Since decoding ToM partially reflects one's ability to interpret mental states from eyes, this task is commonly used to measure an individual's ToM decoding ability (Sabbagh, 2004).

Figure 3.3

Example Stimuli from the Reading the Mind in the Eyes Task



Note. From "The "Reading the Mind in the Eyes" Test revised version: A study with normal adults, and adults with Asperger syndrome or high-functioning autism," by S. Baron-Cohen, S. Wheelwright, J. Hill, Y. Raste, and I. Plumb, 2001, *The Journal of Child Psychology and Psychiatry and Allied Disciplines, 42*(2), p. 242. (https://doi.org/10.1111/1469-7610.00715). Copyright 2001 by Blackwell Publishing. Adapted with permission.

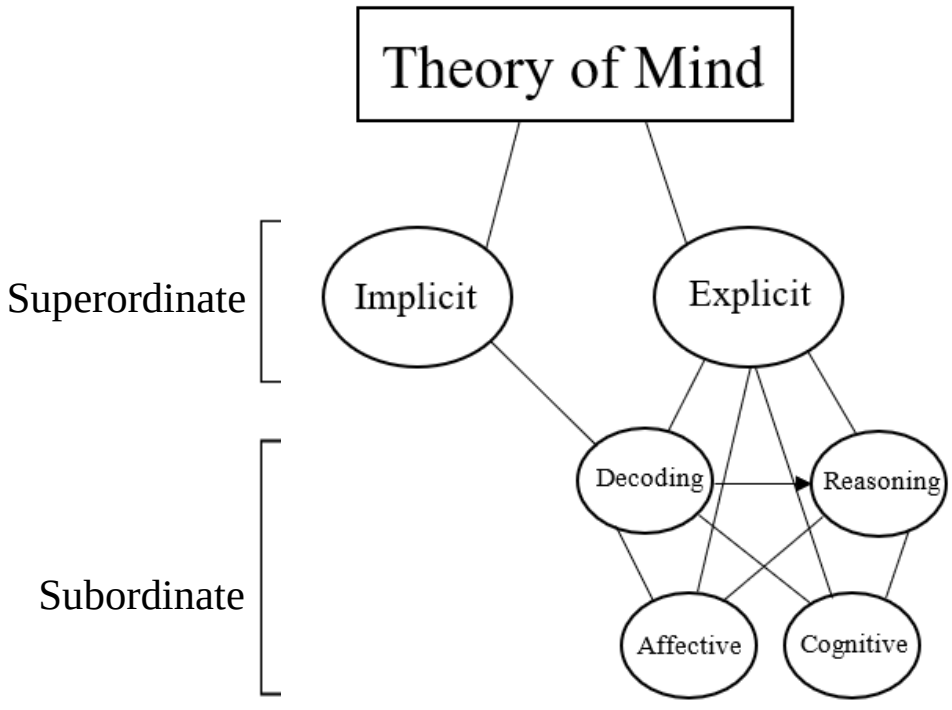
Similarly, tasks initially designed to assess general ToM abilities are commonly employed in measuring reasoning ToM (Bora & Berk, 2016). For example, the Movie for the Assessment of Social Cognition (Dziobek et al., 2006) has individuals view a 15-minute film of four people socializing. This film is split into 46 segments, with each segment ending and asking individuals' a mental state question (e.g., What is X thinking/feeling/intending to do?). Correct responses require individuals to consider and integrate social information (e.g., facial expressions), social context, and interactions between all four characters. Given that this process often requires initial attributions (e.g., how someone is feeling based upon facial expressions) to make inferences about other mental states (e.g., what they might do next), the Movie for the Assessment of Social Cognition plausibly measures reasoning ToM. Thus, the literature supports decoding and reasoning ToM as dissociable and operationalizable subprocesses of explicit ToM. Given that this thesis looks to explore individual differences in

the explicit ToM abilities of neurologically typical adults, employing measures that tap an individual's decoding and reasoning ToM is crucial to ensure an accurate reflection of their true explicit ToM abilities is obtained.

Cognitive and Affective ToM. Cognitive and affective ToM abilities, like decoding and reasoning ToM abilities, are posited to be non-competing subprocesses we are consciously aware of when we make inferences about others' mental states. In addition, cognitive and affective ToM abilities are thought to function in tandem with decoding and reasoning abilities (Figure 3.4; Duclos et al., 2018). Cognitive ToM refers to our ability to understand the beliefs and knowledge of others. Alternatively, affective ToM reflects our ability to understand the feelings of others (Duclos et al., 2018; Shamay-Tsoory et al., 2010). The construct of affective ToM shares substantial conceptual overlap with cognitive empathy. As such, the two terms tend to be used interchangeably throughout the literature (Rogers et al., 2007).

Figure 3.4

The Affective-Cognitive Model of ToM in Relation to the Implicit-Explicit and Decoding-Reasoning Models



Initial evidence for cognitive-affective ToM dichotomy was established through lesion studies. It was observed that ventromedial frontal lobe damage was associated with impaired performance on tasks of affective ToM but not tasks assessing cognitive ToM (Shamay-Tsoory et al., 2003, 2005). Additionally, prefrontal cortex damage is specifically associated with impaired cognitive, but not affective, ToM (Shamay-Tsoory & Aharon-Peretz, 2007). These findings are also supported by functional magnetic resonance imaging data, which show greater activity in the ventromedial prefrontal cortex when completing affective ToM tasks but not cognitive ToM tasks (Sebastian et al., 2012). More recent evidence has further implicated the amygdala in facilitating affective ToM and the parietal lobes in cognitive ToM (Bejanin et al., 2017). Thus, a strong body of evidence supports a distinct and dissociable cognitive-affective ToM system.

Similar to the aforementioned issues with decoding and reasoning ToM, most tasks employed in measuring cognitive and affective ToM were not originally designed and intended to do so. Often, these measures were created to measure ToM *generally* and, as such, indirectly index these subprocesses of ToM. Additionally, measurement tools may simultaneously index decoding/reasoning and cognitive/affective ToM. Exemplifying this, ToM tasks that require individuals to make inferences about the affective states of others (e.g., the Reading the Mind in the Eyes Task) are commonly employed in the measurement of affective ToM (Poletti & Adenzato, 2013; Rominger et al., 2016; Russell et al., 2009). Thereby, performance on the Reading the Mind in the Eyes Task is thought to represent an individual's explicit affective and decoding ToM abilities.

Similarly, measures of reasoning ToM may also index cognitive/affective ToM. The commonly employed measure of reasoning ToM, the Faux Pas Test (Baron-Cohen et al., 1999), requires reasoning about both mental states (cognitive ToM) and emotions (affective ToM). As such, cognitive and affective ToM questions from this measure can be identified

and independently analysed to infer individual cognitive and affective ToM ability (Bottiroli et al., 2016). The Short Story Task, developed to measure reasoning ToM, also appears to index cognitive and affective ToM (Turner & Felisberti, 2017). Thereby, the cognitive-affective distinction offers a conceptually robust and operationalizable theory of ToM that works in parallel with the decoding-reasoning theoretical conceptualisation of ToM. Thus, to index an individual's true explicit ToM abilities, this thesis will explore and employ explicit measures of cognitive-affective and decoding-reasoning ToM to investigate individual differences in the ToM abilities of neurotypical adults.

Criticisms of Existing ToM Theories and Models. Despite attempts within this thesis to provide a coherent model of the intersection between various theoretically dissociable ToM abilities (Figure 3.4, Page 44), it must be acknowledged that there is no universally accepted model, theory, or operationalization, of ToM within the literature (Beaudoin et al., 2020). This is partly due to the number of unique and complex abilities theoretically comprising ToM. Resultingly, several authors have signalled a need for these commonly employed ToM models and theories to be retired to reconstruct a universally agreed-upon structure of ToM (Beaudoin et al., 2020; Schaafsma et al., 2015). In doing so, these authors often attempt to construct frameworks that provide a foundational understanding of the ToM construct from which subsequent literature can be built. For example, Beaudoin et al. (2020) attempted to redefine the ToM construct by grouping existing ToM measures based on their theoretically indexed abilities (e.g., measurement of intentions, emotions, perceptions, and desires). Alternatively, Schaafsma et al. (2015) reconceptualised ToM by defining a basic set of processes that supposedly underpinned ToM (e.g., inferring mental states from actions, inferring mental states from gaze), and these were linked to other cognitive abilities (e.g., understanding of causality).

The stark differences between these two reformulations of ToM highlight the limited agreement about how best to theoretically reconceive this complex construct. As such, neither of these reconstructions or other competing alternatives are widely adopted within the literature. Instead, the aforementioned theories and models (i.e., affective-cognitive, decoding-reasoning, implicit-explicit models) continue to be the dominant theories within the literature. Thus, while the limitations of these ToM models are acknowledged, at the present they offer the most universally utilised model of ToM and the most utility for the present work.

ToM Measurement. Historically, ToM has primarily been measured using false-belief tasks (Brüne, 2003). These tests tap an individual's cognitive and reasoning ToM abilities by testing whether an individual understands that others can hold incorrect information while they (the individual completing the measure) hold correct information (Frith & Corcoran, 1996). Other tests which measure an individual's understanding of social faux pas, attribution of intentions, and deception, have also been utilised to measure a range of ToM abilities (Baron-Cohen et al., 1999; Brunet et al., 2000; Frith & Corcoran, 1996). While commonly used, many of these types of tasks have faced criticism for supposedly poorly operationalising ToM as a construct (François & Rossetti, 2020).

More importantly for this thesis, these tests were designed to assess severe ToM deficits in clinical disorders or to track the development of ToM across childhood. As such, when utilised with neurotypical adults, ceiling effects typically occur (Fitzpatrick et al., 2018; Turner & Felisberti, 2017). Measures that do not show ceiling effects have thereby been created or adopted to measure ToM with this population. The most widely utilised ToM measures with neurotypical adults, which index explicit decoding-reasoning and cognitive-affective ToM, will be reviewed in the following section to inform the measurement of ToM in the present research.

Reading the Mind in the Eyes Test – Revised Edition. The Reading the Mind in the Eyes Test – Revised Edition (RMET) is the most widely employed ToM test with neurotypical adults (Baker et al., 2014). During the test, individuals view 36 images of eyes and must select from a list of four responses the response that best reflects the thought or feeling conveyed by the eyes (Figure 3.3, Page 43; Baron-Cohen et al., 2001). It measures affective (Gallant & Good, 2019) and decoding (Dodell-Feder et al., 2013) ToM abilities. Supporting this, performance on the RMET has been found to correlate positively with measures of other affective abilities (e.g., empathy; Gallant et al., 2020). Supporting the RMET as a measure of decoding abilities, individuals with disorders associated with an impaired ability to accurately interpret perceivable social information (e.g., depression) tend to perform worse on the RMET than healthy controls (Richman & Unoka, 2015). In addition, the RMET displays acceptable levels of convergent validity, evidenced by correlating with theoretically similar constructs, such as recognition of faux pas (r = .28; Ferguson & Austin, 2010). The measure also shows acceptable discriminant validity, by not correlating with dissimilar constructs, such as social desirability (Vellante et al., 2013); and test-retest reliability over one year (*ICC* = .63; Fernández-Abascal et al., 2013). Thus, the RMET is plausibly one adequate measure of decoding and affective ToM.

The RMET is not without limitations, however. Notably, it has been suggested that the RMET is a measure of emotion recognition, as opposed to ToM, given that alexithymia (an inability to identify or describe one's emotional states) is associated with impaired performance (Oakley et al., 2016). This debate is ongoing and unresolved. However, the RMET is still presently widely adopted as a ToM measure. Other criticisms include the measure's lack of ecological validity (Turner & Felisberti, 2017). Specifically, there is no time limit on how long an individual may view each set of eyes. The absence of time limitations plausibly allows individuals to use compensatory strategies, such as vocabulary

knowledge, to determine mental states (Cassels & Birch, 2014). Additionally, impaired performance on the RMET may occur in older neurotypical populations due to the use of static stimuli (Sze et al., 2012). Resultingly, it is recommended that the RMET be used in tandem with cognitive and reasoning ToM measures to adequately capture an individual's true ToM abilities (Turner & Felisberti, 2017). Therefore, the RMET may plausibly be adopted as a measure of affective and decoding ToM, alongside a corresponding reasoning/cognitive measure, within this thesis.

Yoni Test. The Yoni Test was developed as a measure of cognitive and affective ToM and is sensitive to differences in these abilities in neurotypical adults (Kidd & Castano, 2013; Shamay-Tsoory & Aharon-Peretz, 2007). The Yoni Test presents individuals with 96 images of a cartoon face (Yoni). Around Yoni's face are four objects or four other faces with objects next to them. The individual must use the direction of Yoni's eye gaze, Yoni's facial expression, or the gaze and expression of the other four faces to determine mental states. Sixteen control trials do not require mental state inferences to test individual effort.

The Yoni Test allows differentiation of an individual's cognitive and affective ToM abilities. Cognitive trials are based solely on eye gaze, whereas affective trials also require consideration of happy or sad emotive states for Yoni and other faces (Shamay-Tsoory & Aharon-Peretz, 2007). Plausibly, the Yoni Test also requires the use of decoding and reasoning abilities. Second-order trials require the consideration of Yoni's mental state to figure out other characters' mental states, thereby using reasoning. The consideration of eye gaze and emotion invokes the use of decoding abilities. However, as decoding abilities are theoretically employed in the second-order trials, the differential influence of reasoning and decoding abilities on performance cannot be deduced.

There is little evidence to support the psychometric properties of the Yoni Test. The pilot study of the Yoni Test evidenced the measure's construct validity as it correlated with the irony task (r = -.29 to -.35) and the false belief task (r = .25, Shamay-Tsoory & Aharon-Peretz, 2007). There is also some evidence for the discriminant validity of the measure, as it can differentiate between neurotypical individuals and individuals diagnosed with schizophrenia (Shamay-Tsoory et al., 2007). These two studies mark the only psychometric evaluations of this instrument. The measure's validity has also been questioned because of its simplistic stimuli. Specifically, Turner and Felisberti (2017) note that individuals may form basic object-actor associations without engaging in deeper cognitive processes. As such, its appropriateness for use as a measure of cognitive and affective ToM abilities is questionable.

The Movie for the Assessment of Social Cognition. The Movie for the Assessment of Social Cognition (MASC) was designed to assess ToM abilities in individuals with Asperger's and assesses both cognitive-affective and decoding-reasoning ToM (Dziobek et al., 2006; Turner & Felisberti, 2017). The MASC has individuals watch a 15-minute video of four characters interacting at a dinner party. This video is stopped 46 times to ask the individuals to answer questions about characters' thoughts (cognitive ToM), feelings (affective ToM), characters' mental states based upon body language (decoding ToM), and what one character may think of another character based upon contextual information (reasoning ToM). The individual is also asked a series of control questions to ensure they adequately understand the video.

The pilot of the MASC provided initial support for its psychometric properties. Higher scores on the Autism Diagnostic Interview-Revised were associated with poorer performance on the MASC. As autism spectrum disorder is theoretically associated with impairments in ToM, this supported the measure's discriminant validity. The internal consistency (α = .84) and test-retest reliability (ICC = .89 - .92) were also supported (Dziobek

et al., 2006). However, the MASC failed to correlate significantly with the RMET in clinical and non-clinical populations. Only the clinical population's scores on the MASC correlated with the Strange Story Task (Happé, 1994), a ToM measure in which individuals must detect non-literal meanings of character interactions (e.g., sarcasm, irony, white lies) across a series of vignettes. Subsequent studies with similar samples have found contrasting results, observing significant correlations between the RMET and MASC (r = .30 - .51; Fossati et al., 2018; Müller et al., 2016). Thus, despite initial queries about the MASC's validity, subsequent literature has primarily supported the MASC as a valid and reliable measure of a range of ToM abilities.

Short Story Task. The Short Story Task (SST) was initially developed to measure reasoning ToM (Dodell-Feder et al., 2013). The task has individuals read The End of Something by Ernest Hemingway. This story was chosen as Hemingway purposefully omits characters' mental states meaning the reader must infer these themselves. An administrator then reads 14 questions to the examinee, who must verbally respond after each question. These questions assess implicit ToM (1 question), comprehension (5 questions), and ToM reasoning (8 questions). The comprehension questions act as a control task by assessing story understanding.

The SST is also a measure of both cognitive and affective ToM. The examinee is asked to make mental state attributions about the story character's thoughts (cognitive) and feelings (affective; Turner & Felisberti, 2017). However, the marking rubric does not distinguish between cognitive and affective ToM questions. As such, while cognitive and affective ToM abilities may influence performance, the SST cannot be used to differentiate between these ToM abilities.

The SST ToM subscale shows high interrater reliability (*ICC* = .98) and adequate concurrent validity evidenced by correlating with theoretically associated measures (The Interpersonal Reactivity Index Fantasy Subscale and the RMET; *r* = .37-.49; Dodell-Feder et al., 2013). These findings are supported by subsequent replications (Giordano et al., 2019). In addition, it also displays discriminant validity as it can differentiate between neurotypical individuals and individuals diagnosed with schizophrenia in remission (Fekete et al., 2020). However, it displays poor internal consistency, likely due to varying levels of item difficulty and the wide range of mental states assessed (Dodell-Feder et al., 2013; Giordano et al., 2019). Given these psychometric properties, the SST appears to be an adequate measure of reasoning ToM with neurotypical adults while also reflecting affective and cognitive ToM.

Measurement of ToM for the Current Thesis. Chapter 1 identified a gap within our understanding of the relationship between video game play and ToM abilities in neurotypical adults. This gap was to be initially investigated through a correlational investigation of the relationship between ToM and video game genre engagement. This chapter has highlighted that should this investigation prove fruitful and a follow-up pretest-post-test experimental design was to be conducted, existing measures are currently inadequate due to the absence of alternate forms. Thus, in selecting the primary measure of ToM in this thesis, the construction of an alternate form had to be possible. This thesis, therefore, aimed to develop an alternate form of a ToM measure that can be used with neurologically typical adults.

The Yoni test presented the greatest number of limitations for use. While alternate forms may plausibly be constructed, the absence of adequate psychometric evaluations (Shamay-Tsoory et al., 2007), in tandem with the limitations outlined above by Turner and Felisberti (2017), raised concerns about the validity of this measure. Alternatively, the RMET had extensive support for its psychometric properties (F. J. Ferguson & Austin, 2010; Fernández-Abascal et al., 2013; Gallant et al., 2020; Vellante et al., 2013). However, the

RMET presented an issue in its theoretical measurement of ToM and the feasibility of alternate form construction. As a primary measure of ToM in this thesis, the RMET posed the issue of only measuring affective and decoding ToM, thereby requiring the parallel use of another ToM measure. Additionally, the creation of 36 equivalent stimuli, each measuring 36 different affective states, posed practical limitations given the scope of this thesis as a Doctorate of Clinical Psychology thesis. Thus, the Yoni Test and the RMET were not adopted as primary measures of ToM.

Theoretically, the MASC is a more robust measure of ToM compared to the SST. While both have adequate and similar psychometric properties, the MASC allows for the assessment of both cognitive-affective and decoding-reasoning ToM. However, practical considerations presented limitations to its use for the present research. Creating a video with sufficiently complex character interactions that allowed for the construction of 46 equivalent questions posed a substantial barrier to adoption. Similarly, measurement procedures of the MASC are extremely time and resource intensive as the large number of times the video is interrupted means that the testing time for each participant would need to be very high. Alternatively, creating an alternate form of the SST required selecting a new story and constructing only eight equivalent ToM questions. Further, it can easily be adapted to a written or online delivery format, requiring comparatively little assessment time and resource use. In light of these considerations, the SST is selected as the primary measure of ToM adopted for the present thesis. Thereby, an alternative form of the SST was developed, and its psychometric properties were investigated.

CHAPTER 4:

OVERVIEW OF THE CURRENT THESIS

As noted earlier in this literature review, the effect of video games on cognition has been extensively studied within various subdomains. Yet, there is an absence of research examining the effects of video game play on Theory of Mind (ToM), a broad and commonly utilised socio-cognitive ability. It is unclear what effect engagement with different video game genres or social contexts may have on these abilities. As such, the first aim of this thesis was to determine whether engagement with different video game genres or social contexts was associated with differences in the ToM abilities of neurotypical adults. Study 1 explored this through correlational analyses with data collected via an online survey. It was anticipated that Study 2 would examine the presence of causal relationships between these variables through a pretest-post-test experimental design, should analyses in Study 1 prove fruitful.

Existing theory within the video game literature has primarily been adopted to predict how violent video game play influences aggression. Given that research examining relationships between social cognition and video game play is a new and emerging area, existing theories have not been extended to account for relationships between these variables. Study 1 explored whether relationships between video game genre engagement and differences in ToM were better explained by the General Aggression Model (GAM) or the General Learning Model (GLM). Study 1 also examined whether the relationship between the social context of video game play and ToM abilities was consistent with predictions made GLM. Initially, it was anticipated that Study 2 would extend these findings and determine whether the predictions made by the GLM or GAM were more accurate in predicting causal changes in ToM abilities using an experimental paradigm. To foreshadow, unexpected

findings in Study 1, combined with issues surrounding difficulties undertaking research during Covid-19 restrictions in New Zealand, altered the trajectory of these research plans.

Currently, there are no measures of ToM for use with neurotypical adults that have alternate forms for use in a pretest-post-test experimental design. Thereby, repeated use of the same measure meant prior familiarity with the measure in a pre-test might bias post-test scores. Resultingly, Study 1 also aimed to develop and pilot an alternate form of the Short Story Task (SST) for use in Study 2. Thus, the following primary and secondary aims were anticipated to be explored through two studies in this thesis:

Study 1 aims:

1. To develop and pilot an alternate version of the Short Story Task with neurotypical adults.
2. To determine whether engagement with different video game genres is associated with differences in the ToM abilities of neurotypical adults.
    2.1. To determine whether the social context of video game play is related to differences in an individual's ToM abilities.
3. To determine whether the GLM or GAM proposes a more accurate theoretical framework for investigating the influence of video game play on ToM abilities.
    3.1. To determine whether relationships between ToM ability and the social context of video games are consistent with predictions made by the GLM.

Study 2 aims (as originally planned):

1. To provide experimental evidence toward causal inferences about whether engagement with a specific video game genre (identified in Study 1) is associated with differences in the ToM abilities of neurotypical adults.
2. To determine whether the GLM or GAM proposes a more accurate theoretical framework for investigating the causal changes in ToM abilities as a result of video game play.

However, subsequent findings from Study 1 - in tandem with lockdown restrictions imposed due to the Covid-19 pandemic - suggested that an experimental paradigm for Study 2 was unlikely to be fruitful or feasible. The results of Study 1 suggested that the Short Story Task B may be an improved measure of ToM relative to the SST, a finding that required replication in a subsequent study. Study 1 also highlighted that while research examining links between ToM ability and video games is scant and mixed (Bormann & Greitemeyer, 2015; Kühn et al., 2019), an area of growing interest is in links between reading literary fiction and ToM improvement. While initial evidence has suggested familiarity with literary fiction is associated with improvements in ToM (Mumper & Gerrig, 2017), this body of literature is small. In addition, replication of findings within the psychological literature is paramount, given the ongoing replication crisis (Open Science Collaboration, 2015). As a result, an alternative online survey was conducted for Study 2 to investigate the following aims:

Study 2 aims (alternative):

1. To determine whether the Short Story Task or the alternate form piloted during Study 1 is a better measure of ToM with neurotypical adults.

1.1. To determine whether self-report scales of ToM ability and short story comprehension are adequate substitutes for tests of ToM ability and short story comprehension.

2. Replicate the finding that familiarity with literary fiction is associated with improvements in ToM abilities (Mumper & Gerrig, 2017).

CHAPTER 5:

STUDY 1 METHOD

In this chapter, I will discuss the methods for Study 1. The methods will be discussed in two stages. In the first stage, I will discuss the creation of the Short Story Task B (SST-B) and the video game play questionnaire. In the second stage, I will discuss the pilot testing process for the SST-B and data collection procedures.

Stage 1 – Creation of the SST-B and Video Game Play Questionnaire

Story Selection. Story selection was restricted to works created by Ernest Hemingway. As the Short Story Task (SST) uses the story *The End of Something* (Hemingway, 2003), another Hemingway story was selected to ensure that differences in prose between stories did not account for differences in participant scores. Hemingway also purposefully omits characters' mental states throughout most of his writing. Thus, participants must invoke their ToM abilities to deduce characters' thoughts, intentions, beliefs, and emotions throughout his stories (Dodell-Feder et al., 2013).

Additional criteria were also imposed upon story selection. Namely, Dodell-Feder et al. (2013) used the Flesch Reading Ease Score (FRES; Flesch, 1948) and the Flesch-Kincaid Grade Level (Kincaid et al., 1975; FKGL) to evidence the readability of the story. The FRES produces a score between 0 – 100 (scores closer to 100 indicate that a story is easier to read) based on the total number of syllables, words, and sentences within a story (Appendix A). The FKGL also considers the total number of syllables, words, and sentences within a story but produces a score indicating the United States grade level to which a story should be interpretable (Appendix A). *The End of Something* has a FRES of 92.7 and a FKGL of 2.8. Therefore, the story adopted in the SST-B needed similar FRES and FKGL scores.

Finally, story length was considered in selecting the story for the SST-B. *The End of Something* is 1427 words in length. A story of similar length was deemed necessary to ensure adequate complexity of character interactions upon which questions could be generated. If a story were too short, these interactions would likely lack the necessary complexity to assess Theory of Mind (ToM). If the story was too long, fatigue or increased demands on ToM might have resulted in score differences between the two measures. Given these considerations, four criteria were outlined in consideration of story selection:

1. Ernest Hemingway must write the story.
2. The story must have sufficiently complex character interactions.
3. The story must produce similar FRES and FKGL scores to *The End of Something.*
4. The story should be of a similar length to *The End of Something*.

The selection of appropriate texts was restricted to The Complete Short Stories of Ernest Hemingway (Hemingway, 1987) to accommodate length considerations. Stories were then read and categorised based on character interactions. This process resulted in the selection of two potential texts: *A Simple Enquiry* (Hemingway, 1927a) and *Ten Indians* (Hemingway, 1927b). *A Simple Enquiry* (989 words) centres on a Major and his subordinate in the army. Across the story, the Major implicitly questions his subordinate about his sexuality with the implication of them potentially beginning a sexual relationship. Characters are described as displaying a range of non-verbal communication methods and showing emotions such as shame and embarrassment. Alternatively, *Ten Indians* (1587 words) centres on a boy who has attended a Fourth of July celebration but is then informed by his father that his girlfriend has been unfaithful. Again, in this story, characters are described as showing a range of non-verbal communication methods, empathy, and emotions such as sadness, disgust, and pity. However, *Ten Indians* was deemed inappropriate for modern audiences,

given the presence of racist and sexist undertones across the story (Schedler, 2013). These were not present in *A Simple Enquiry*. Thus, readability analyses were not conducted for *Ten Indians*.

Readability scores were calculated for *A Simple Enquiry* (FRES = 91.5, FKGL = 3.1). These scores were similar to those observed for *The End of Something*. As such, *A Simple Enquiry* met all four outlined criteria and was deemed an appropriate story for use in the construction of the SST-B.

Question Construction. Question design was modelled on procedures outlined by Dodell-Feder et al. (2013). Specifically, five questions were designed to assess participants' story comprehension and eight were designed to assess reader ToM abilities. The implicit ToM question was omitted due to concern about the validity of implicit ToM measures (Kulke et al., 2018, 2019; Poulin-Dubois & Yott, 2018).

Where possible, questions for the SST-B were designed to follow similar wording and structure to the SST to improve the likelihood that potential differences between participant scores resulted from differences in understanding of mental state content as opposed to question comprehension. How questions on the SST match questions on the SST-B is shown in Table 5.1. This process was not possible for all questions due to differences in the plot and character interactions across the stories. To partially rectify this, the ending of *A Simple Enquiry* was edited to more closely resemble the ending of *The End of Something*. These edits allowed the construction of a parallel version of the SST questions 7 and 11. Edits did not significantly impact story readability (FRES = 90.4, FKGL = 3.2) or story length (968 words).

No parallel version of the SST questions 6 and 10 could be plausibly constructed. SST question 6 is a context-specific question about a character quote with an implicit meaning.

Similarly, question 10 is a context-specific question about a character's actions and emotions following an argument. Creating a parallel question for the SST-B would have required significant revisions to *A Simple Enquiry*. These revisions would have impacted the overall narrative, significantly deviating from Hemingway's original work. This deviation may have, in turn, led to participants incorrectly interpreting the original character's interactions, thoughts, and emotions in the context of the required changes. As such, an additional variation of the SST question 8, and a unique mental state question (question 9), were constructed for the SST-B.

Table 5.1

Comparison of SST Questions to Equivalent Question Versions on the SST-B

SST Questions	SST-B Questions
1. What do Nick and Marjorie observe on the shoreline as they are rowing to the point to set their fishing lines?	**1.** What can be observed outside the hut's windows?
2. What does Nick mean when he says, "They aren't striking?"	**4.** What does the adjutant mean when he says, "Be soft, Pinin … The major is sleeping"?
3. Nick and Marjorie have a pail of perch for what purpose?	**2.** The major has a saucer of oil for what purpose?
4. Do Marjorie's actions suggest that she is experienced or inexperienced at fishing? What makes you say that?	**3.** Do the adjutant's actions suggest he is hardworking or lazy? What makes you think that?
5. Why does Nick say to Marjorie, "You know everything"?	**6.** Why does the major say "Tonani … can you hear me talking?"
6. Why does Marjorie reply, "Oh Nick, please cut it out! Please, please don't be that way!"?	No equivalent question.
7. Why is Nick afraid to look at Marjorie?	**12.** Why does Pinin avoid eye contact with James?

Table 5.1 (Continued)

	7. Why does the major say “All right … You needn’t be superior.”?
8. What does Nick mean when he says, “It isn’t fun anymore”?	**5.** What does Pinin mean when he replies, “I have been with girls.”
9. Why does Marjorie sit with her back toward Nick when she asks, “Isn’t love any fun?”?	**8.** Why does Pinin look at the floor when the major asks him “And you really don’t want-” and “That your great desire isn’t really-”?
10. Why does Marjorie take the boat and leave and what is she feeling at that moment?	No equivalent question.
11. Who is Bill and what does he reveal when he asks Nick, “Did she go alright? … Have a scene?”?	**11.** Who is James and what does he reveal when he asks, “So the major propositioned you too?”
12. What is Nick feeling when he says, “Oh, go away, Bill! Go away for a while”?	**10.** What is Pinin feeling when he leaves the major’s room and walks outside?
13. The story is called “The End of Something.” What is the title referring to?	**13.** The story is called “A Simple Enquiry.” What is the title referring to?
No equivalent question.	**9.** Why is the major ‘really relieved’?

Marking Rubric Construction. Construction of the marking rubric was developed based on criteria outlined by Dodell-Feder et al. (2013). For the comprehension subscale, a score of 0 indicated an entirely incorrect response; a score of 1 indicated a partially correct response; a score of 2 indicated an entirely correct response. The marking rubric was designed so that most individuals could perform at the ceiling if they appropriately engaged with the story. For example, the opening line of *A Simple Enquiry* is, “Outside, the snow was higher than the window.” (Hemingway, 1927a, p. 107). The corresponding comprehension question was, “What can be observed outside the huts window?”. Participants received full marks if they stated, “Snow; snow that is higher than the window; melted snow; snow melted into a trench.” Similar guidelines were used to assign scores of 0, 1, and 2 to responses for

the ToM subscale. Additionally, higher scores were achieved if a participant considered multiple characters' mental states, identified characters' underlying intentions, identified characters' emotions, and correctly interpreted body language. Therefore, scores may range from 0 – 10 for the comprehension subscale and 0 – 16 for the ToM subscale. The SST and SST-B questions and marking rubrics are observable in Appendices B-1 and B-2.

Literary analyses of *A Simple Enquiry* were used to ensure a correct answer on the marking rubric reflected Hemingway's intended emotion/thought/intention for the character (Charles Jr, 1995). To achieve full marks for question 12, based on my edits to the story, participants had to indicate why a character may be avoiding eye contact. The correct answer was derived from literature informing common body language when one experiences embarrassment/shame (Edelmann & Hampson, 1979; Modigliani, 1971). Finally, correct answers for question 11 (Appendix B-2), which were also constructed based on story edits, were derived from the correct answers for the equivalent question on the SST.

Video Game Play Questionnaire. No widely used measure exists to examine how participants engage with different video game genres. Thus, the primary purpose of the video game play questionnaire was to examine how participants differed on the three dimensions of video games commonly implicated in cognitive changes (time, content, context; Gentile, 2011). To measure time spent playing video games, participants were asked to indicate their weekly video game play time on a scale from 0 – 30+. The upper limit of 30 hours was chosen arbitrarily.

To examine the social context of participants' video game play, participants were asked whether they commonly played single-player, multi-player, or both types. Whether a game is played cooperatively or competitively can be associated with different cognitive or behavioural changes (Ewoldsen et al., 2012). Thus, if participants indicated that they played

multi-player games, they would then be asked if these were more commonly cooperative or competitive. While Gentile's (2011) conceptualisation of video game context is broader than the social context, alternative contextual factors were deemed too genre specific to be relevant to each participant. Therefore, these were not examined.

Video game content was operationalised by asking participants to rank 14 genres of video games from most played to least played. An additional information sheet was attached to the video game questionnaire that defined each genre, with associated examples of video games, to allow informed choices to be made by participants when ranking. Currently, there is no consensus regarding what video game genres are valid or how they may be defined (Blocker et al., 2014). Therefore, this list of 14 genres was derived from those outlined by Lenhart et al. (2008), with associated definitions being informed by Lenhart et al. (2008) and Adams (2014). Genres, and a list of definitions are shown in Table 5.2.

Table 5.2

Video Game Genres and Associated Definitions

Genre	Definition
Sports	A sports video game simulates some aspect of an athletic sport.
Role-Playing	A role-play game is one in which the player controls character(s) and guides them through a series of quests.
First Person Shooter	Video games in which the player uses guns, or other weapons, to shoot targets from a first-person perspective.
Action	Video games in which the challenges presented are tests of player's physical skills and coordination. Puzzle solving, tactical conflict, and exploration challenges are often present.
Racing	Video games which mimic the experience of driving a car.

Table 5.2 (Continued)

Adventure	Video games that involve an interactive story about a protagonist character who is controlled by the player.
Strategy	Video games in which challenges presented are strategic conflicts.
Simulator	Video games that simulate real or imaginary situations. This may include activities such as piloting vehicles or controlling characters artificial lives.
Puzzle	Video games where puzzle solving is the primary activity.
Rhythm	Video games which challenge a player's sense of rhythm. This may involve dancing or playing an instrument.
Fighting	Video games which simulate hand to hand combat.
Survival Horror	Video games where the character is vulnerable and under armed. They commonly involve a scary setting, a scarcity of gun ammunition, awkward camera angles, and puzzle solving.
Massive Multiplayer Online Game (MMOG)	Online spaces where multiple individuals play a videogame together. In game activities may be played solo or in a group. In MMOGs, the world and gameplay continue to move forward even when an individual or group of players is not playing the game.
Virtual World	Video games that simulate environments or spaces where users, represented by avatars, communicate or interact simultaneously or synchronically.

Stage 2 – Pilot of the SST-B and Examining Relationships Between Video Game Engagement and ToM

Participants. The sample size was calculated using G*Power. An a priori power analysis, set at .90 power to detect a medium effect size of r = .30 at the .05 alpha error probability, indicated that a sample size of 112 participants was required. An r = .30 was selected in line with Cohen's (1988) effect size conventions to detect a 'medium' effect size. A medium effect size was anticipated, given that experimental effect sizes reported by Bormann and Greitemeyer (2015) between video game play and affective ToM were medium.

One hundred and twelve participants ranged in age from 18 to 67, with the mean age being 29 years. Males (70%) accounted for the majority of participants, while females (27%), non-binary individuals (3%), and those who preferred not to say (1%) accounted for the remaining. While the high proportion of males in the sample is curious, it may partially reflect the sex split of video game players, whereby males are more likely to be video game players (Entertainment Software Association, 2020). Individuals came from various educational backgrounds, with completion of upper secondary school (32%) or a bachelor's degree (33%) being the most common. Further demographic information is observable in Table 5.3.

Table 5.3

Demographic Characteristics of Participants in Study 1

Characteristic	Frequency (N=112)	Percent (%)
Age (Years), Mean (SD)	29.3 ± 11.4	
Gender		
Male	78	69.6
Female	30	26.8
Non-Binary	3	2.7
Prefer not to say	1	0.9
Education		
Early Childhood	1	0.9
Primary	5	4.5
Lower Secondary	1	0.9
Upper Secondary	36	32.1
Post-Secondary (Non-Tertiary)	18	16.1
Short Cycle Tertiary	7	6.3
Bachelor's Degree	37	33.0
Master's Degree	5	4.5
Doctoral Degree	2	1.8

Note. SD = Standard Deviation

Inclusion and Exclusion Criteria. Study objectives examined how playing different video games may be related to ToM. Thus, active video game play was required to conduct these analyses. Participants were required to indicate their weekly video game playtime on the video game questionnaire. Initially, I preregistered that a minimum of one hour of weekly video game gameplay would be acceptable. However, Prolific, the survey distribution platform, only allowed participants to indicate their weekly video game playtime in multiples

of 3 (e.g., 0-3 hours weekly, 3-6 hours weekly, 6-9 hours weekly, 10-12 hours weekly, 12+ hours weekly). Therefore, a minimum of 3 hours was implemented due to limitations in Prolific's screening questions. The minimum number of reported hours was four.

Participants were required to be from a western country as ToM is a culturally bound construct (Oi et al., 2013). Therefore, ToM measures are more likely invalid for individuals from non-western English-speaking countries, given these measures are often developed, piloted, and used with western populations. As such, only individuals from New Zealand, Australia, Canada, the United States, or the United Kingdom were recruited to increase the likelihood that the employed ToM measures would be valid. Participants were required to complete the survey on a desktop. This restriction was employed due to the implementation of the RMET for online use. Completion required access to a keypad to select a response from one through to four. Additionally, the images were designed for viewing on a desktop computer. Completing the RMET on a phone or other device may have led to changes in image resolution, potentially impacting participant responses.

Participants were required to be aged 16 or above. This restriction was imposed to comply with Massey University's code of ethical conduct regarding recruiting children for research. Additionally, the SST and SST-B were designed for adult populations, and it is unclear if they would be appropriate for use with younger populations. Participants were also required not to have a severe visual impairment. This restriction was to ensure that all participants would be able to perceive visual stimuli on the RMET adequately. Finally, participants were required to not be currently or previously diagnosed with any neurological, developmental, or psychological disorders. This final criterion was employed as ToM is found to be impaired in most of these disorders (Cotter et al., 2018). To mitigate this potential confound, these individuals were excluded.

Thus, to be eligible to take part in the present study, participants were required to:

1. Be actively playing video games for a minimum of three hours per week.
2. Be currently living in New Zealand, Australia, Canada, the United States, or the United Kingdom.
3. Complete the survey on a desktop computer.
4. Be over the age of 16.
5. Not have a severe visual impairment.
6. Have no current or previous diagnoses of neurological, developmental, or psychological disorders.

Participant Recruitment. All participants were recruited online using Prolific. Prolific is an online survey-hosting website where potential participants are informed about studies they may self-select to partake in. Outlined inclusion criteria were implemented using Prolific's custom pre-screening tool. Participants were only informed about the study if they completed studies on a desktop; were from New Zealand, Australia, Canada, the United States, or the United Kingdom; and played video games for a minimum of 3 hours a week. Participants were redirected to an information sheet (Appendix C-1) if they expressed interest. If they consented to participate, they were prompted to click to the next page of the survey.

Measures. In addition to the SST-B and video game play questionnaire, various measures were selected to examine the outlined research objectives. These measures and the rationale for their selection will be discussed.

Autism-Spectrum Quotient. The Autism-Spectrum Quotient (AQ) is a 50-item self-report measure of traits associated with the autistic spectrum (Baron-Cohen et al., 2001). In this study, the AQ was used to assess the predictive validity of the SST and SST-B. The AQ

presents participants with 50 statements who must choose one of four response options for each statement: definitely agree, slightly agree, slightly disagree, and definitely disagree. The measure has five subscales that assess social skills, attention switching, attention to detail, communication, and imagination. A slight or definite endorsement of a trait associated with autism receives a score of one. The measure has a total score of 50, with a score above 32 indicating the possible presence of autism spectrum disorder. The measure displays face validity, construct validity (α = .63-.77), and test-retest reliability (r = .82 - .92; Baron-Cohen et al., 2001; Stevenson & Hart, 2017). Importantly, clinical and subclinical scores on the AQ have been shown to negatively correlate with scores on common ToM tests (Baron-Cohen et al., 2001; Mintah & Parlow, 2018). In the present study, internal consistency reliability for the AQ was α = .81.

In this study, Cronbach and Meehl (1955) informed how predictive validity was defined. Specifically, they define predictive validity as the correlation between two variables when scores on one test are obtained sometime after the first test. As scores on the AQ were collected after scores on the SST/SST-B, this technically reflects predictive validity. Larger periods are often adopted between testing periods in predictive validity analyses, however for the purposes of the present thesis I consider this shortened period between the instruments a proof-of-concept test for the presence of short-term predictive validity.

Reading the Mind in the Eyes Test - Revised Edition. The RMET is one of the most widely used ToM tests (Baker et al., 2014). Given this, this measure was used to assess the concurrent validity of the SST and SST-B. During the test, participants view 36 images of eyes and must select from a list of four responses the response that best reflects what they believe the individual is thinking/feeling (Baron-Cohen et al., 2001). The RMET is commonly conceptualised as a measure of affective (Gallant & Good, 2019) and decoding (Dodell-Feder et al., 2013) ToM abilities. The RMET displays acceptable levels of

convergent validity, evidenced by correlating with theoretically similar constructs, such as recognition of faux pas (r = .28; Ferguson & Austin, 2010); discriminant validity, by not correlating with dissimilar constructs, such as social desirability (Vellante et al., 2013); and test-retest reliability over one year (*ICC* = .63; Fernández-Abascal et al., 2013). Notably, in the context of this study's aims, the RMET has also been shown to positively correlate with the scores on the SST ToM subscale (r = .35 - .49; Dodell-Feder et al., 2013; Giordano et al., 2019).

However, the RMET commonly displays low internal consistency, which may reflect more than one underlying factor (Olderbak et al., 2015). In the present study, internal consistency reliability for the RMET was α = .71. While alternative structures have been proposed (Olderbak et al., 2015; Preti et al., 2017), these are not widely used. The current study adapted the RMET from its original pencil and paper delivery format to an online format. Participants were presented with an image and pressed the 1, 2, 3, or 4 keys to select the associated mental state they believed was depicted. Similar delivery methods of the RMET have been successfully utilised in prior research (Panero et al., 2016).

Short Story Task. The SST plausibly reflects reasoning, cognitive, and affective ToM (Dodell-Feder et al., 2013; Turner & Felisberti, 2017). As the SST should theoretically be an alternate form of the SST-B, it was used to assess the alternate form's reliability of the SST-B. The SST has the participant read The End of Something by Ernest Hemingway. An administrator then reads 14 questions to the examinee, who must verbally respond after each question. These questions assess implicit ToM (1 question), comprehension (5 questions), and cognitive and affective ToM reasoning (8 questions). The SST ToM subscale shows high interrater reliability (*ICC* = .98) and adequate concurrent validity, evidenced by correlations with theoretically associated measures (r = .37-.49; Dodell-Feder et al., 2013). However, it displays poor internal consistency, likely due to varying item difficulty levels and the wide

range of mental states assessed. The present study adapted the test from verbal delivery and response to a written format. The implicit ToM question was omitted due to conflicting evidence on the validity of implicit ToM as a dissociable ability and queries regarding current assessment methods (Kulke et al., 2018, 2019; Poulin-Dubois & Yott, 2018). Additionally, the screening questions “Did you read it for school or pleasure?”, “What grade were you in?”, “What class was it for?”, “Is the story familiar to you?” and “Do you know anything about the story? What do you know about it?” were omitted as they were deemed repetitive and unnecessary and therefore unduly burdensome on participants’ time.

Procedure. After consenting to participation, participants were prompted to enter their Prolific ID to match response sets to associated Prolific submissions. This process ensured that only participants who completed the study received monetary compensation. Participants were randomly selected to complete either the SST or SST-B first. All participants completed both the SST and SST-B. Measure presentation was counterbalanced across participants to account for practice and order effects. After reading the instructions and story, participants were asked, “Have you read this story before?” If ‘Yes’ were selected, participants would then be asked: “How long ago did you read it?”, “What do you remember about the story?” and “Have you discussed the story with anyone?”. If ‘No’ was selected, these questions did not appear. All participants indicated they had not read either story previously. The subsequent 13 comprehension and ToM questions were presented on the same page as the story so participants could refer back to it if necessary.

All participants were then presented with the video game questionnaire to reduce potential practice effects that completion of the SST/SST-B would have on the later presented counterbalanced measure by maximizing time between measure presentations. This allowed for the maximum amount of memory interference and decay between the two measures. Keeping this questionnaire in the same position standardized this timeframe across

participants. An additional attention check question was added at the beginning of the video game questionnaire.

The RMET was presented in the third position for all participants. It was hypothesised that completing the SST/SST-B may improve performance on the RMET due to the effects of reading literary fiction on ToM (Dodell-Feder & Tamir, 2018). By standardising the presentation, this effect was consistent across all participants. The RMET was prefaced with a task description and a list of mental state descriptors that would be shown across the task. Participants could then select those they did not know and receive an associated definition. Upon task completion, participants were asked on a ten-point scale how difficult they found the task, how much mental effort it required, and how well they thought they did. They were also asked to indicate if they had completed this task before.

Participants were then presented with the SST or SST-B, depending on which measure was completed earlier in the study. Following this, participants completed the AQ. The AQ was presented last as it has a high degree of face validity regarding its measurement of traits of autism spectrum disorder. Autism spectrum disorder may be associated with deficits in ToM abilities (Baron-Cohen et al., 2001). Therefore, to mitigate demand effects on participants' ToM abilities, it was completed after all ToM measures. A second attention check was embedded after question 33 on the AQ.

Participants were then asked their age, gender, level of education, and whether they were proficient English speakers. One participant indicated they did not have proficiency in English. This individual was included in the final analyses as their scores on all ToM measures were within the preregistered acceptable range (± 3.29 standard deviations; Tabachnick et al., 2007). Following this, participants were asked whether they encountered any issues with the survey and to guess what the study hypotheses may be. This was to check

whether potential awareness of study aims had impacted or biased responses. If participants correctly guessed study aims, their data was examined to determine whether this had resulted in deviations from expected performance across all measures (i.e., whether their scores across measures were ± 3.29 standard deviations). If so, their data for that measure was to be omitted. Finally, participants were provided with a debrief section explaining the study's aims and whether they would like to be informed of the study's results later.

Preregistered Analysis Strategy. All statistical analyses, excluding the psychometric evaluation of the SST, were preregistered on the Open Science Framework (10.17605/OSF.IO/CS6BA). All statistical analyses were conducted using SPSS version 25.0. Data were initially screened for missing data using Little's (1988) Missing Completely At Random (MCAR) test. As missing data was MCAR, the expectation-maximization algorithm was used. Data was also inspected for normality and the presence of outliers. Outliers were defined as ± 3.29 standard deviations from the mean (Tabachnick et al., 2007). Skewness and kurtosis were calculated for all measures to determine the appropriateness of parametric tests, assess normality, and determine the presence of ceiling effects. Spearman and Pearson's correlations were calculated between the SST and SST-B ToM and respective comprehension subscales to determine whether story comprehension was related to performance on the ToM subscale.

Statistical analyses for the psychometrics properties of the SST and SST-B primarily followed the procedures outlined (Dodell-Feder et al., 2013). For interrater reliability analyses, all forms were marked by the primary researcher. Then, 25% of the forms were randomly selected to be marked by another independent marker. The second marker was an employee of the International Media Psychology Laboratory at Massey University and held postgraduate tertiary qualifications. The marker was provided with anonymised participants' responses, the SST and SST-B stories, and the marking rubrics. No training on scoring was

provided. Intra-class Correlation Coefficients (ICC) estimates, based on an absolute agreement model, were then calculated independently for the comprehension and ToM subscales. These were interpreted in reference to guidelines outlined by Koo and Li (2016; Table 5.4). Estimates were not compared to values obtained by Dodell-Feder et al. for the SST, as the form of ICC used was not reported.

Table 5.4

Koo and Li's Benchmark Scale for ICC's

ICC Statistic	Strength
< .50	Poor
0.50 – 0.75	Moderate
0.75 – 0.90	Good
> .90	Excellent

Internal consistency reliability for the SST and SST-B was measured using Cronbach's alpha. Values were calculated independently for the comprehension and ToM subscales. Cronbach's alpha for the SST-B was determined as acceptable if $\alpha = .45$ or greater for the ToM subscale and $\alpha = .25$ or greater for the comprehension subscale. While these were below the generally accepted convention of $\alpha = .70$, Dodell-Feder et al. (2013) observed Cronbach's alpha values of $\alpha = .54$ (ToM subscale) and $\alpha = .31$ (comprehension subscale) on the SST due to the range of content and varying question difficulties within these subscales.

Alternate forms reliability for the SST-B was measured using both ICC's, based upon a consistency model. ICC estimates were interpreted in reference to the guidelines outlined by Koo and Li (2016; Table 5.4, Page 75).

Concurrent validity for the SST and SST-B was measured using Spearman's rank order correlations between the SST and SST-B ToM subscales and the RMET. Spearman correlations were used in place of Pearson correlations as the RMET data was significantly skewed (Z score = -3.64, $p = .0003$). The standard alpha value of .05 was used to determine whether the SST and SST-B ToM subscales displayed concurrent validity with the RMET if $p < .05$. Dodell-Feder et al. (2013) found a medium effect size of $r = .49$ between the RMET and SST. As such, concurrent validity was determined to be acceptable if $r \geq .30$. Spearman correlations were also calculated between the SST and SST-B comprehension subscales and the RMET to determine whether the comprehension questions assessed mental state abilities.

Predictive validity for SST and SST-B was measured by calculating Pearson correlations between the AQ and the SST and SST-B ToM subscales. These values were then squared to determine how much variability in AQ scores was explained by performance on these measures. While the AQ is not a diagnostic tool, a score above 32 or above on the AQ may indicate the presence of autism spectrum disorder, which may be associated with ToM impairments (Baron-Cohen et al., 2001). Analyses with and without individuals who scored above this threshold were conducted to analyse whether these scores affected the SST and SST-B's predictive validity. The standard alpha value of .05 was used to determine whether the SST and SST-B ToM subscale predicted scores on the AQ. Additionally, predictive validity was anticipated to be low, as sub-clinical AQ scores commonly show small negative correlations with ToM measures (e.g., r = -.21; Mintah & Parlow, 2018). Given this, predictive validity was deemed acceptable if scores on the SST-B ToM subscale accounted for ≥ 1% of the variation in AQ scores (a correlation of approximately $r = .10$).

To determine whether playing different video game genres was related to ToM ability, data was initially reverse coded (genres ranked 1 received a score of 14; genres ranked 14 received a score of 1). Spearman's rank order correlations were then used to

independently calculate correlations between engagement with all 14 video game genres and performance on the SST. To determine whether video game genre engagement was related to SST ToM subscale scores, the standard alpha value of .05 was used.

Finally, a one-way between-subjects ANOVA was used to determine whether participant scores of the SST differed depending on whether they played single-player, multi-player, or both. To determine whether engagement with these different gaming modalities was related to SST ToM subscale scores, the standard alpha value of .05 was used. Additional exploratory analyses were also conducted and are explicitly identified as exploratory in the results chapter of this thesis.

CHAPTER 6:

STUDY 1 RESULTS

Data Screening

Subsequent statistical procedures and interpretive guidelines were preregistered on the Open Science Framework (https://doi.org/10.17605/OSF.IO/87QKX). The Short Story Task (SST) psychometric evaluation was not preregistered due to experimenter error. Thereby, they may be considered exploratory and require replication. Further deviations from this preregistration and exploratory analyses are explicitly outlined.

Statistical Outliers. Total and appropriate subscale scores for all measures were analysed for univariate outliers. In line with current convention, outliers were deemed to be scores that fell outside 3.29 Standard Deviations of the mean (SD; Leys et al., 2019; Tabachnick et al., 2007)

No outliers were detected for the Autism-Spectrum Quotient (AQ) full-scale score. Similarly, no outliers were detected on the Short Story Task B (SST-B) comprehension and ToM subscales. One outlier was observed on the Reading the Mind in the Eyes Test (RMET). This data point was removed. Additionally, one individual completed the RMET on a mobile device. This data point was removed as the test was constructed for completion on a desktop only. Finally, one individual completed the RMET twice. Scores for the second completion were removed, while the first completion scores were retained.

Two participants' scores were excluded on the ToM subscales of the SST and SST-B as they performed poorly on the comprehension subscales. The comprehension subscale plausibly functions as a manipulation check, as individuals are anticipated to perform at or near ceiling (Dodell-Feder et al., 2013). If individuals perform poorly on this subscale, it is unlikely they adequately understand the story. As such, they likely have not engaged their

ToM abilities, rendering their resulting data practically uninterpretable. This data exclusion criterion was not preregistered.

Missing Data. Every item was analysed for missing values. Two data points (1.8%) were missing for the question, “How many hours a week do you play video games?”. Little’s (1988) Missing Completely at Random test indicated that these data points were most likely missing completely at random. Given that these missing data points represented less than 5% of the overall item scores, the expectation-maximization algorithm was used to impute data in SPSS version 25 to improve statistical power (Scheffer, 2002).

Fifteen people (14.3%) failed to answer any items on the RMET. Participants were asked at the survey’s conclusion whether they had experienced any issues. One participant commented that they did not complete the RMET as “I accidentally skipped it by pressing the wrong button”. Another commented, “In the section where I was asked to rate the emotion shown in a person's eyes, none of the questions showed up on the screen after the tutorial.” Thus, this missing data is hypothesised to be due to survey design or incompatibility with some participants desktops. As all items for the RMET were missing for these individuals, they were excluded from analyses involving the RMET.

Skewness, Kurtosis, and Test Selection. The application of parametric tests requires data to follow a normal distribution. Therefore, the symmetry and pointedness of the data were calculated using z-scores based on skewness and kurtosis (Table 6.1). In line with conventions outlined by Kim (2013), a z-score of 3.29 or greater was used to evidence a departure from normality (as n = 50-300). Given this, non-parametric tests were utilised to analyse the SST Comprehension Subscale and the RMET.

Notably, there was substantial variation in the scores across both measures (Figure 6.1). Across both measures, no individuals received scores of 14, 15, or 16, indicating an absence of ceiling effects.

Table 6.1

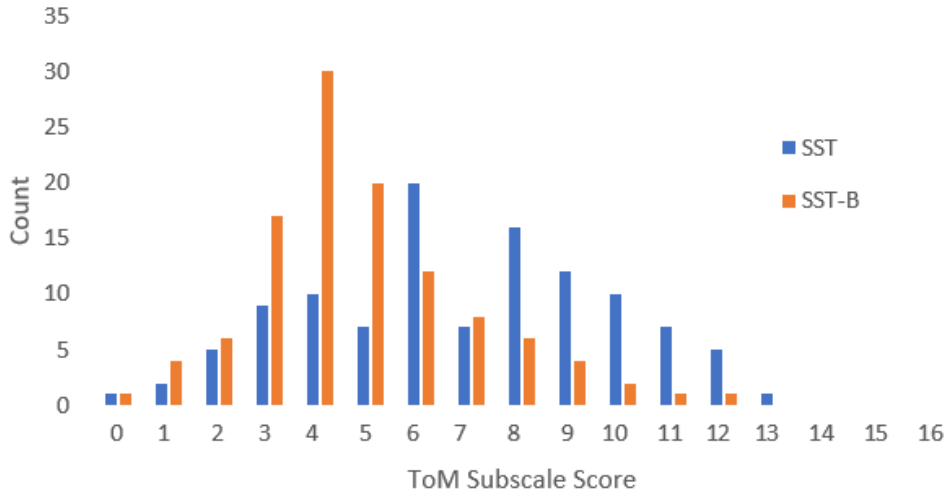
Skewness and Kurtosis Indices for Measures and Subscales

Measures	n	Skewness			Kurtosis		
		Statistic	z-score	p	Statistic	z-score	p
SST							
Comprehension	110	-1.26	-5.48	<.001	1.44	3.15	.002
Theory of Mind	110	-0.05	-0.20	.84	-0.78	1.70	.089
SST-B							
Comprehension	110	-0.51	-2.25	.024	-0.64	1.40	.16
Theory of Mind	110	0.74	3.24	.001	0.80	1.77	.077
RMET	95	-0.90	-3.64	.0003	1.28	2.61	.009
AQ	112	0.49	2.14	.032	0.32	0.71	.48

Note. Standard error of the mean for skewness and kurtosis were .23 and .45 for the SST-B and AQ respectively; .23 and .46 for the SST respectively; and .25 and .49 for the RMET respectively.

Figure 6.1

Distribution of the ToM Subscale Scores for the SST and SST-B.



Practice Effects. Before conducting analyses involving the RMET, a Mann-Whitney U test was conducted to determine whether self-reported familiarity was the RMET was associated with improvements in test performance. As can be seen from Table 6.2, test familiarity was not associated with improved performance on the RMET.

Table 6.2

Mann-Whitney U Results Comparing Previous Completion of the RMET and RMET Total Score

Familiarity	n	Mean Rank	U	p
Familiar	18	48.8	630.5	.55
Unfamiliar	77	44.5		

Story Comprehension and ToM Subscale Performance. Performance on the comprehension subscales of the SST and SST-B were correlated against their respective ToM subscales. This was to determine whether story comprehension was related to performance on the ToM subscale. Spearman correlation coefficients are observable in Table 6.3. Across both measures, significant positive correlations emerged between ToM and their respective comprehension subscales across both the SST and SST-B. However, correlations were stronger for the SST-B than the SST. These findings suggest that improved performance on one subscale is related to improved performance on the other.

Table 6.3

Spearman Correlations Between the SST and SST-B Comprehension Subscales and their Respective ToM Subscales

Measure	Correlation (95% CI)	p
SST	.25 (.07, .42)	.008
SST-B	.37 (.20 , .52)	<.001

Reliability Evaluation

Inter-Rater Reliability. All inter-rater reliability estimates were based upon a single measure, absolute agreement, two-way random effects model. Intraclass Correlation Coefficients (ICC) and associated 95% confidence intervals are observable in Table 6.4. Across both subscales, the SST-B displayed greater ICC estimates than the SST. For the SST comprehension and ToM subscale, estimates fell in the 'good' and 'moderate' range, respectively. For the SST-B comprehension and ToM subscale, estimates fell in the 'excellent' and 'good' range, respectively

Table 6.4

Inter-Rater Reliability Estimates for the Total Scale Scores of the SST and SST-B Comprehension and ToM Subscales

Measure	ICC	95% CI	p
SST			
Comprehension	.75	.53 , .88	<.001
Theory of Mind	.57	-.02 , .82	<.001
SST-B			
Comprehension	.92	.83 , .96	<.001
Theory of Mind	.75	.10 , .92	<.001

Note. CI = Confidence Interval.

Internal Consistency Reliability. Internal consistency for the SST and SST-B was calculated using Cronbach's alpha. As outlined previously, Cronbach's alpha was determined to be acceptable if $\alpha = .45$ or greater for the ToM subscale and $\alpha = .25$ or greater for the comprehension subscale. While this is below the generally accepted convention of $\alpha = .70$ (Gliem & Gliem, 2003), it is similar to the results of Dodell-Feder et al. (2013) who observed

Cronbach's alpha values of α = .54 (ToM subscale) and α = .31 (comprehension subscale) on the SST due to the range of content and varying question difficulties within these subscales.

Table 6.5 presents the alpha coefficients for the SST and SST-B. Cronbach's alpha for both SST-B subscales exceeded the minimum outlined acceptable criterion. In contrast, while the SST ToM subscale fell within the acceptable range, the comprehension scale fell below the minimum pre-registered criterion for acceptable internal reliability. As Cronbach's alpha reflects the intercorrelations between test items, and correlation size may be reduced when data lack variability (Goodwin & Leech, 2006), the observed alpha is anticipated to be the result of low variance in scores across these items. Supporting this, increasing item score variability through the inclusion of outliers resulted in a Cronbach's alpha of .46.

Table 6.5

Cronbach's Alpha for the SST and SST-B

Measure	α (95% CI)
SST	
Comprehension	.23 (-.03 , .43)
Theory of Mind	.61 (.49 , .71)
SST-B	
Comprehension	.48 (.31 , .62)
Theory of Mind	.56 (.43 , .67)

Note. CI = Confidence Interval. Number of items in ToM subscale = 8; Number of items in comprehension subscale = 5.

Alternate Forms Reliability. Alternate forms reliability was estimated using a multiple measure, consistency, two-way random effects ICC model. An ICC value ≥ .50 was preregistered as acceptable. In retrospect this criterion was lax given evidence that higher

ICCs are preferable As such, ICCs will also be interpreted in reference to the criteria outlined by Koo and Li (2016).

Table 6.6 displays the ICC between the SST and SST-B. Conventional measurement of parallel forms reliability assumes the resultant measures will have equal means and variances and thus utilises an absolute agreement model of ICC. As equal means were not anticipated, consistency ICC was adopted. Given this deviation from standard application of the ICC, it should be interpreted with caution. Thus, while the observed ICC value of .59 falls above the preregistered acceptable value of .50 and indicates moderate reliability (Koo & Li, 2016), this remains lower than what would be anticipated of true alternate forms.

Table 6.6

ICC between the SST and SST-B

	Correlation (95% CI)	F	p
ICC	.59 (.40 , .72)	2.418	<.001

Note. CI = Confidence Interval. Two-tailed.

Validity Evaluation

Concurrent Validity. Spearman correlations between the RMET and the SST and SST-B ToM subscales were calculated to determine the concurrent validity of the SST and SST-B. To determine whether performance on the RMET was related to story comprehension, Spearman correlations were calculated between the SST and SST-B comprehension subscale scores and the RMET. Concurrent validity was deemed acceptable for the SST and SST-B if $r_s \geq .30$, given that Dodell-Feder et al. (2013) observed a medium effect size of $r_s = .49$ between the RMET and SST.

Spearman correlations are observable in Table 6.7. Consistent with hypotheses, Spearman correlations indicated significant convergence between the SST and SST-B ToM subscales and the RMET. Additionally, the correlation between the SST-B ToM subscale and the RMET meet the acceptable threshold of $r_s \geq .30$. Contrastingly, the SST did not meet the acceptable threshold as $r_s < .30$. Further complicating the picture, the SST-B comprehension subscale significantly correlated with the RMET while the SSTs did not. This correlation suggests the SST-B comprehension subscale may inadvertently be indexing some aspects of ToM rather than purely comprehension. While the correlation between the RMET and SST comprehension subscale was non-significant, this may again reflect the lack of variation in scores on this subscale. Additionally, inadequate power may be limiting the analysis' ability detect this weak correlation.

Table 6.7

Spearman Correlations Between the SST and SST-B Subscales and the RMET

Measure	Correlation (95% CI)	p
SST		
Comprehension	.18 (-.02 , .37)	.09
Theory of Mind	.24 (.04 , .42)	.02
SST-B		
Comprehension	.37 (.18 , .53)	<.001
Theory of Mind	.46 (.29 , .61)	<.001

Note. CI = Confidence Interval. Two-tailed.

Predictive Validity. Pearson correlations between the AQ and the SST and SST-B ToM subscale were calculated and squared to determine the predictive validity of the SST

and SST-B. Spearman correlations were calculated between the RMET and AQ as a reference, given that the RMET has been shown to negatively correlate with scores on the AQ (Eyuboglu et al., 2018). As previously mentioned, a score of 32 or above on the Autism Spectrum Quotient (AQ) may indicate the presence of autism spectrum disorder, which may in turn be associated with ToM impairment (Baron-Cohen et al., 2001). Thereby, an analysis was conducted with and without individuals who scored above this threshold to analyse whether these scores affected the predictive validity.

Predictive validity was anticipated to be low, as sub-clinical scores on the AQ commonly only show small negative correlations with ToM measures (e.g., r = -.21; Mintah & Parlow, 2018). Given this, predictive validity was pre-registered to be acceptable if scores on the SST-B ToM subscale accounted for ≥ 1% of the variation in AQ scores.

Results are observable in Table 6.8. When including scores above the clinical threshold, the SST had a non-significant but acceptable $R^2 = .01$. This suggests that participants' SST scores account for 1% of the variation in scores on the AQ. However, contrasting the hypothesised outcome, the SST-B and RMET showed non-significant positive correlations with the AQ. With the clinical threshold included, $R^2 = .02$ for the SST-B, indicating the SST-B explained 2% of the variance in AQ scores in the direction inverse to what was anticipated. Upon removal of individuals' scores above the AQ clinical threshold, the negative trend for the SST reversed, and a non-significant positive correlation was observed. Stronger positive correlations between the AQ, RMET, and SST-B were also observed, but these were also non-significant. $R^2 = .03$ and .04 for the SST and SST-B, respectively.

Table 6.8

Pearson and Spearman Correlations between the SST ToM Subscale, SST-B ToM Subscale, RMET and the AQ

Measure	Clinical Threshold Included				Clinical Threshold Excluded			
	n	Correlation (95% CI)	R^2	p	n	Correlation (95% CI)	R^2	p
SST	110	-.10 (-.28, .09)	.01	.31	94	.16 (-.03 , .34)	.03	.12
SST-B	110	.14 (-.05 , .32)	.02	.15	96	.19 (-.01 , .38)	.04	.06
RMET	95	.04 (-.17 , .24)	.00	.72	79	.07 (-.15 , .29)	.00	.53

Note. CI = Confidence Interval. The AQ threshold for clinical significance is ≤ 32. SST and SST-B correlations were calculated using Pearson correlations. RMET correlations were calculated using Spearman correlations.

Video game Engagement and ToM

Video game Genre Engagement and ToM Ability. In line with preregistered criteria, Spearman's rank order correlations were used to independently calculate correlations between engagement with all 14 video game genres and performance on the SST ToM subscale (Table 6.9). It was hypothesised that strategy games, role-playing games, virtual world games, and massive online multi-player games would significantly positively correlate with ToM abilities. Correlations ranged in strength from r_s = -.11 to r_s = .12. Contradicting the hypotheses for this study, all correlations were non-significant, suggesting that playing different video game genres is unrelated to performance on the SST ToM subscale.

Table 6.9

Spearman Correlations Between the SST and Rankings of Engagement with Video game Genres

Video game Genre	Correlation	p
Virtual Worlds	-.11	.26
MMOGs	.10	.30
Survival Horror	.12	.23
Fighting	-.02	.81
Rhythm	-.10	.30
Puzzle	-.09	.33
Simulator	.01	.89
Strategy	.02	.84
Adventure	-.04	.71
Racing	.03	.76
Action	.02	.87
FPS	-.09	.34
Role Playing	.12	.20
Sports	.01	.93

Note. Two-tailed. Scores were reverse coded such that genres ranked 1 were assigned a score of 14 while genres ranked 14 were assigned a score of 1.

Multi-Player Versus Single-Player Video Games. Table 6.10 shows a one-way between-groups ANOVA was used to determine whether engagement with multi-player, single-player or playing both video game types an equal amount was related to performance on the SST ToM subscale. It was hypothesised that playing multi-player video games, or playing both an equal amount, would be associated with higher scores on this task compared

to those who played only single-player games (i.e., Multi-Player > I play both an equal amount > Single-Player).

Table 6.10

One-Way Between Subject ANOVA Comparing Multi-Player, Single-Player, or 'I Play Both an Equal Amount' Group Effects for the SST ToM Subscale

Measure	n	M	SD	F	p	df	Effect Size (η^2)
Multi-Player	18	7.78	2.92	0.875	.42	2, 107	0.02
Plays Both	50	6.86	2.89				
Single-Player	42	6.76	2.77				

Note. M = Mean, SD = Standard Deviation

Levene's test was initially carried out to determine whether the assumption of homogeneity of variances was met, $F(2, 107) = .07$, $p = .93$. Thus, the assumption was met, and the application of parametric tests was appropriate. The ANOVA did not reveal a statistically significant difference for performance on the SST ToM subscale between groups ($p > .05$). Given this, no post hoc analyses were undertaken. These results contradict the hypotheses, suggesting that playing single or multi-player games is unrelated to performance on the SST ToM subscale.

Exploratory Analyses

Subsequent analyses were not preregistered and are exploratory. Their primary purpose was to generate future hypotheses. Interpretation of such analyses will be conducted with consideration for this limitation.

RMET and Video Game Engagement. The SST is purported to measure cognitive, affective, and reasoning ToM abilities (Dodell-Feder et al., 2013). Alternatively, the RMET theoretically measures affective and decoding ToM abilities (Dodell-Feder et al., 2013; Gallant & Good, 2019). Given that the RMET theoretically measures different underlying abilities of ToM, Spearman correlations were calculated between the RMET and engagement with different video game genres. These analyses were to determine whether these ToM abilities were potentially related to engagement with different video game genres.

Outputs are observable in Table 6.11. Correlations ranged in strength from r_s = -.26 to r_s = .17. A significant negative relationship emerged between RMET performance and fighting game engagement (r_s = -.26, p = .01). Similarly, a significant negative relationship was observed between performance on the RMET and engagement with action games (r_s = -.24, p = .02). These results suggest that higher self-reported engagement with fighting and action video games is associated with slightly poorer performance on the RMET.

Table 6.11

Spearman Correlations Between the RMET and Rankings of Engagement with Video game Genres

Video game Genre	Correlation	p
Virtual Worlds	0	.97
MMOGs	.09	.41
Survival Horror	.03	.78
Fighting	-.26	.01
Rhythm	.10	.10
Puzzle	-.02	.82
Simulator	.04	.73
Strategy	-.03	.79
Adventure	-.04	.73
Racing	.07	.49
Action	-.24	.02
FPS	-.06	.55
Role Playing	.17	.11
Sports	.13	.21

Note. Two-tailed. Scores were reverse coded such that genres ranked 1 were assigned a score of 14 while genres ranked 14 were assigned a score of 1.

SST-B and Video Game Engagement. While the SST-B should theoretically measure the same underlying ToM abilities as the SST, preregistered analyses suggested that the SST-B may have greater validity than the SST. As the SST-B may provide a comparative or better reflection of an individual's ToM abilities within this study, as compared to the SST, Spearman correlations between SST-B ToM subscale scores and self-reported engagement with different video game genres were undertaken.

Results are shown in Table 6.12. Correlations ranged in strength from r_s = -.18 to r_s = .16. No significant results were observed, indicating that engagement with different video game genres was not related to performance on the SST-B. These results contrast those observed in Table 6.11 when correlating the RMET with video game genre engagement. Specifically, both action and fighting games did not significantly correlate with the SST-B ToM subscale, despite the relatively strong correlation between SST-B ToM subscale scores and fighting game engagement.

Table 6.12

Spearman Correlations Between the SST-B ToM Subscale and Rankings of Engagement with Video game Genres

Video game Genre	Correlation	p
Virtual Worlds	0	.99
MMOGs	-.08	.41
Survival Horror	-.06	.50
Fighting	-.18	.053
Rhythm	.13	.18
Puzzle	-.01	.95
Simulator	.06	.55
Strategy	0	.99
Adventure	0	.99
Racing	.12	.23
Action	-.08	.38
FPS	-.11	.24
Role Playing	.16	.09
Sports	.07	.48

Note. Two-tailed.

SST-B and Video Game Engagement Controlling for Weekly Video game Play.

The amount of video game play can influence cognition, such that the more an individual plays a game, the greater the magnitude of associated cognitive change (Gentile, 2011). Thus, spearman correlations were calculated between the SST-B ToM subscale scores and engagement with different video game genres while controlling for weekly hours of video game play to determine whether time playing video games influenced the interactions between these measures.

Results are observable in Table 6.13. Correlations ranged in strength from r_s = -.19 to r_s = .16. This is compared to r_s = -.18 to r_s = .16 when weekly video game play was not controlled. This resulted in a significant correlation between engagement with fighting games and performance on the SST-B (r_s = -.19; p = .049). These results suggest that higher self-reported engagement with fighting video games might be associated with poorer performance on the SST-B ToM subscale when controlling for weekly hours of video game play time, however the exploratory nature of these analyses suggest that these analyses should be interpreted with caution.

Table 6.13

Spearman Correlations Between the SST-B and Rankings of Engagement with Video game Genres Controlling for Weekly Hours of Gameplay

Video game Genre	Correlation	p
Virtual Worlds	0	.99
MMOGs	-.08	.42
Survival Horror	-.06	.50
Fighting	-.19	.049
Rhythm	.13	.18
Puzzle	-.01	.95
Simulator	.06	.55
Strategy	0	.99
Adventure	0	.99
Racing	.12	.23
Action	-.09	.38
FPS	-.11	.23
Role Playing	.16	.09
Sports	.07	.48

Note. Two-tailed

Multi-Player Versus Single-Player Video Games Using the RMET and SST-B.

Tables 6.14 and 6.15 show a Kruskal-Wallis H Test and a one-way between groups ANOVA that were used to determine whether engagement with multi-player, single-player or playing both video game types an equal amount was related to performance on the RMET or SST-B ToM subscale. The Kruskal-Wallis H Test was adopted over a one-way ANOVA for the RMET due to data being non-normal.

Table 6.14

Kruskal-Wallis H Test Comparing Multi-Player, Single-Player, or 'I Play Both an Equal Amount' Group Effects for the RMET and SST-B ToM Subscale

Measure	n	M	SD	MR	H	p	df
Multi-Player	15	27.67	2.92	50.40	.256	.88	2
Plays Both	43	26.95	4.52	48.53			
Single-Player	37	26.78	4.31	46.41			

Note. M = Mean, SD = Standard Deviation, MR = Mean Rank

Table 6.15

One-Way Between Subject ANOVA Comparing Multi-Player, Single-Player, or 'I Play Both an Equal Amount' Group Effects for the SST ToM Subscale

Measure	n	M	SD	F	p	df	Effect Size (η^2)
Multi-Player	18	5.11	2.25	.328	.72	2, 109	.01
Plays Both	51	4.67	2.33				
Single-Player	43	4.93	2.04				

Note. M = Mean, SD = Standard Deviation

Both analysis strategies did not reveal a statistically significant difference based on group membership ($p > .05$). Given this, no post hoc analyses were undertaken. These results align with earlier findings utilising the SST ToM subscale and suggest that playing single or multi-player games is unrelated to performance on the RMET or SST-B ToM subscale.

SST and SST-B Exploratory Factor Analyses. Exploratory factor analyses were conducted for the full-scale SST and SST-B combined and separately. The primary purpose of these analyses was to observe the underlying factor structure of these measures and how items across the two subscales loaded onto these underlying latent factors. It was initially anticipated that as both the SST and SST-B have two subscales each, two latent factors would emerge and be retained. Items on the ToM subscales were anticipated to load onto factor one, while items on the comprehension subscales would load onto factor two.

All factor analytic procedures used principal axis factoring and no rotation, as how each item loaded onto each factor was of interest. Only factors with eigenvalues greater than one were retained for initial analyses. Further, only items with factor loadings ≥ .30 were retained as items that accounted for < 9% of explained variance were not of interest

Outputs for combined factor analysis of the SST and SST-B are observable in Table 6.16. The primary purpose of this analysis was to observe whether the SST and SST-B subscales were loading onto the same two latent factors. Questions 5 and 12 on the SST ToM subscale and 9 and 10 on the SST-B ToM subscale did not have a factor loading ≥ .30 on the first factor. All other ToM questions across both measures had factor loadings ≥ .30, with the SST question 7 having the highest factor loading of .587.

Interestingly, question 1 on the SST comprehension subscale, and questions 2, 3, 4, and 13 on the SST-B comprehension subscale, exhibited factor loadings ≥ .30 on factor one. This finding aligns with the results in Table 6.7, whereby the SST-B comprehension questions appeared to be partially indexing ToM abilities rather than purely assessing comprehension. Subsequently, none of the SST-B's comprehension questions had a factor loading ≥ .30 on the second factor. Further, only question 13 on the SST comprehension subscale had a factor loading ≥ .30 on the second factor. This suggests that the SST's

comprehension subscale may also be inadequately indexing story understanding. Alternatively, it may reflect the heterogeneity of the questions' content. These results suggest the retention of only a single factor, which is hypothesized to reflect the underlying latent factor of ToM. This conclusion is supported by Figure 6.2, whereby the point of inflection on the scree plot indicates that only factor one should be retained.

Independent exploratory factor analyses were then conducted for the SST and SST-B. The primary purpose of these analyses was to observe the extent to which these measures were independently indexing an underlying latent factor.

Table 6.16

Factor Matrix for SST and SST-B Comprehension and ToM Questions [a]

Question	Factor								
	1	2	3	4	5	6	7	8	9
SST-B Comprehension Q1						-.365			
SST-B Comprehension Q2	.398								
SST-B Comprehension Q3	.337		.323						
SST-B Comprehension Q4	.479		.523						
SST-B ToM Q5	.406	-.409						.376	
SST-B ToM Q6	.459	-.344							
SST-B ToM Q7	.362	-.580							
SST-B ToM Q8	.359					-.497			
SST-B ToM Q9									
SST-B ToM Q10									
SST-B ToM Q11	.306								
SST-B ToM Q12	.363						.324		
SST-B Comprehension Q13	.336			.398	.381				
SST Comprehension Q1	.425								
SST Comprehension Q2				-.399					
SST Comprehension Q3			.405						
SST Comprehension Q4			-.301						
SST ToM Q5									
SST ToM Q6	.318				.378				-.354
SST ToM Q7	.587								
SST ToM Q8	.411	.388							
SST ToM Q9	.405								

Table 6.16 (Continued)

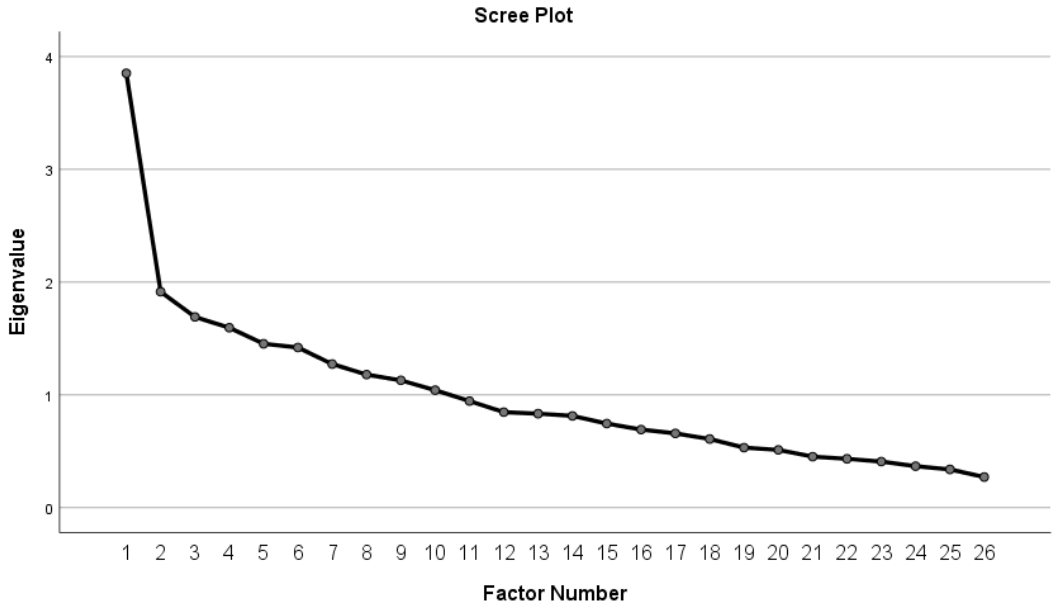
SST ToM Q10	.459		.385	-.413
SST ToM Q11	.459	-.397		
SST ToM Q12				
SST Comprehension Q13	.352			

Note. Extraction Method: Principal Axis Factoring.
Loadings <.3 not reported

a. Attempted to extract 10 factors. More than 25 iterations required. (Convergence=.003).

Figure 6.2

Scree Plot Displaying Eigenvalues for the Combined SST and SST-B Comprehension and ToM Subscales



Outputs for independent exploratory factor analyses of the full-scale SST are observable in Table 6.17. In line with the combined factor analysis, ToM items 5 and 12 did not have factor loadings ≥ .30 on the first factor, while items 6 to 11 did. Question 6 had the minimum acceptable loading of .309, while question 7 had the maximum loading of .710. Comprehension items 1 and 13 had loadings ≥ .30 on factor one. However, only comprehension item 3 had a factor loading ≥ .30 on factor two. These results suggest a similar pattern to that observed in Table 6.16.

Table 6.17

Factor Matrix for the SST [a]

Questions	Factor				
	1	2	3	4	5
SST Comprehension Q1	.345				
SST Comprehension Q2			.429		
SST Comprehension Q3		.556	.497		
SST Comprehension Q4			-.303	.493	
SST ToM Q5					
SST ToM Q6	.309				
SST ToM Q7	.710				-.322
SST ToM Q8	.459				
SST ToM Q9	.513				
SST ToM Q10	.359	.436		-.407	
SST ToM Q11	.509				
SST ToM Q12					.442
SST Comprehension Q13	.325				

Note. Extraction Method: Principal Axis Factoring.
Loadings <.3 not reported.
a. Attempted to extract 5 factors. More than 25 iterations required. (Convergence=.008). Extraction was terminated.

Outputs for independent factor analysis of the full-scale SST-B are observable in Table 6.18. In line with results observable in Table 6.16, ToM question 10 did not have a factor loading ≥ .30. However, contrasting the combined factor analysis, comprehension

question 3 and ToM question 11 did not have a factor loading ≥ .30, while ToM question 9 did. Question 6 had the minimum loading still considered within the acceptable range of .339, while question 7 had the maximum loading of .619. Additionally, comprehension questions 2, 4, and 13 had factor loadings ≥ .30 on factor one, while online question 3 had a loading ≥ .30 on factor two. Again in line with the independent factor analysis for the SST, these results suggest a similar pattern to that observed in Table 6.16.

Table 6.18

Factor Matrix for the SST-B [a]

Questions	Factor					
	1	2	3	4	5	6
SST-B Comprehension Q1					.417	
SST-B Comprehension Q2	.411				.305	
SST-B Comprehension Q3		.470	.465			
SST-B Comprehension Q4	.393		.371			
SST-B ToM Q5	.469			.353		
SST-B ToM Q6	.483					
SST-B ToM Q7	.619	-.502			-.371	
SST-B ToM Q8	.492		-.465	-.370		
SST-B ToM Q9	.339					
SST-B ToM Q10						.364
SST-B ToM Q11						
SST-B ToM Q12	.331	.441	-.319	.319		
SST-B Comprehension Q13	.364					

Note. Extraction Method: Principal Axis Factoring.
Loadings <.3 not reported.
a. Attempted to extract 6 factors. More than 25 iterations required. (Convergence=.008).

Following full scale exploratory factor analyses, exploratory factor analyses were conducted combined and independently on only the ToM subscales of the SST and SST-B to observe item loadings without interference from the comprehension items.

Outputs for combined factor analysis of the SST and SST-B ToM subscales are observable in Table 6.19. In line with the combined full-scale factor analysis (Table 6.16), SST ToM question 5 and 12, and SST-B ToM questions 9 and 10, did not exhibit factor loadings ≥ .30. SST-B question 8 had the minimum loading still in the acceptable range of .314, while SST question 7 had the maximum loading of .625. These results suggest some degree of common variance between the measures across the ToM subscales.

Table 6.19

Factor Matrix for the SST and SST-B ToM Questions [a]

Questions	Factor						
	1	2	3	4	5	6	7
SST-B ToM Q5	.430	-.347					
SST-B ToM Q6	.541	-.316		-.391			
SST-B ToM Q7	.371	-.560					
SST-B ToM Q8	.314		.303		-.487		
SST-B ToM Q9		-.332					
SST-B ToM Q10						.359	
SST-B ToM Q11	.365				.413		
SST-B ToM Q12	.393		.337				
SST ToM Q5							
SST ToM Q6	.321		-.462				
SST ToM Q7	.625					-.427	
SST ToM Q8	.405	.371		-.386			
SST ToM Q9	.389						
SST ToM Q10	.453						
SST ToM Q11	.528						.377
SST ToM Q12							

Note. Extraction Method: Principal Axis Factoring.
Loadings <.3 not reported.
a. Attempted to extract 7 factors. More than 25 iterations required. (Convergence=.003).

Outputs for the independent factor analysis of the SST ToM subscale are observable in Table 6.20. In line with previous results, questions 5 and 12 had a factor loading <.30 for

the first factor. Alternatively, question 12 had a factor loading of .551 for factor two. These findings provide some support for this measure having an underlying coherent factor, a finding that is consistent with the output of the combined ToM subscale factor analysis.

Table 6.20

Factor Matrix for the SST ToM Questions [a]

Question	Factor		
	1	2	3
SST_Q5			
SST_Q6	.348	-.307	
SST_Q7	.711		
SST_Q8	.391		
SST_Q9	.509		
SST_Q10	.375		-.362
SST_Q11	.528		
SST_Q12		.551	

Note. Extraction Method: Principal Axis Factoring.
Loadings <.3 not reported.
a. Attempted to extract 3 factors. More than 25 iterations required. (Convergence=.003). Extraction was terminated.

Outputs for the independent factor analysis of the SST-B ToM subscale are observable in Table 6.21. Contrasting earlier findings in Table 6.19, independent analyses resulted in question 9 having a factor loading ≥ .30 for factor one. However, questions 11 and 12 had stronger factor loadings for factor two as opposed to factor one, and question 8 had a stronger factor loading for factor three as opposed to factor one. However, as most questions had loadings ≥ .30 on factor one, this also suggests an underlying coherent factor is present for the ToM subscale.

Table 6.21

Factor Matrix for the SST-B ToM Questions [a]

Question	Factor			
	1	2	3	4
SSTB_Q5	.627		-.464	
SSTB_Q6	.552			
SSTB_Q7	.566	-.316		
SSTB_Q8	.355		.362	
SSTB_Q9	.374			
SSTB_Q10				
SSTB_Q11	.313	.481		.343
SSTB_Q12	.334	.553		

Note. Extraction Method: Principal Axis Factoring.
Loadings <.3 not reported.

a. Attempted to extract 4 factors. More than 25 iterations required. (Convergence=.004). Extraction was terminated.

Validity Analysis Using Factor Scores. Factor scores were calculated for the SST, SST-B, and combined SST and SST-B ToM subscales. These factor scores were calculated to rerun earlier validity analyses so as to further investigate the validity of these measures. Notably, correlating these factor scores against the RMET may provide evidence to suggest what the underling latent construct for factor one is, given that the RMET theoretically measures ToM. Thus, factor scores for the SST ToM subscale, SST-B ToM subscale, and combined SST and SST-B ToM subscales, were calculated and correlated against the RMET.

Results are observable in Table 6.22. All factor one factor scores across the SST, SST-B, and combined ToM subscales significantly correlated in the expected direction with the RMET. In line with initial correlations between the SST, SST-B, and RMET (Table 6.7),

the correlation between the SST-B factor one factor scores and the RMET was considerably stronger than that observed between the SST and RMET. While the combined SST and SST-B factor scores had the strongest correlation with the RMET of .45, the SST-B factor scores alone were comparable with an effect size of .44. These results suggest the SST-B alone is likely doing a comparatively better job of indexing ToM, as measured by the RMET, than the original SST.

Table 6.22

Spearman Correlations Between the SST ToM Subscale, SST-B Subscale, and Combined SST and SST-B ToM Subscales Factor Scores and the RMET

Measure	Correlation	p
SST	.26	.01
SST-B	.44	<.001
Combined	.45	<.001

Note. Two-tailed. Factor one factor scores used. Factor scores calculated using Bartlett's method.

CHAPTER 7:

STUDY 1 DISCUSSION

Currently, the measurement of Theory of Mind (ToM) abilities in neurologically typical adults is difficult due to a lack of validated behavioural measures for this population (Turner & Felisberti, 2017). One measure that shows promise for measuring ToM in neurologically typical adults is the Short Story Task (SST; Dodell-Feder et al., 2013). However, only a single version of this measure exists. As a result, practice effects may bias scores when using this measure in pretest-post-test experimental designs, ABA designs, longitudinal studies, or with people familiar with the story.

The first aim of this study was to assess the psychometric properties of the Short Story Task B (SST-B), a newly constructed alternate form of the SST. Additionally, only two studies have investigated how video game play may influence ToM abilities (Bormann & Greitemeyer, 2015; Kühn et al., 2019). These studies did not investigate how video games' overarching content and context may be implicated in these changes (Gentile, 2011). Therefore, the second aim of this study was to examine how engagement with different video game content and contexts was associated with ToM abilities. Finally, it is unclear whether relationships between video game content and context and differences in ToM were better explained by the General Aggression Model (GAM) or the General Learning Model (GLM). The third aim of this study was to determine whether the GLM or GAM proposes a more accurate theoretical framework for investigating the influence of video game content and context on ToM abilities.

Examination of the Psychometric Properties of the SST and SST-B

A primary aim of this study was to examine the reliability and validity of the SST-B to determine whether this measure may reflect ToM. In addition, I examined whether there was support for the use of the SST-B as an alternate form of the SST. Concurrently, the

psychometric properties of the SST were also examined to determine, within the context of this study, whether it was a valid measure of ToM.

Validity. Preregistered concurrent and predictive validity analyses were conducted to examine the validity of the SST and SST-B within this study, in addition to non-preregistered exploratory factor analyses. Regarding the predictive validity of these measures, findings from the current study partially supported the predictive validity of the SST. Participant scores on the SST were non-significantly negatively correlated with scores on the Autism Spectrum Quotient (AQ) with R^2 values indicating that as little as 1% of variance was explained by the association. Although the size of this association exceeded pre-registered criterion for acceptable validity, it was non-significant. However, in contrast, the SST-B had a non-significant positive correlation with the AQ, thereby suggesting that the predictive validity of the SST-B was not supported. Interestingly, the RMET also had a non-significant positive correlation with the AQ, indicating that this measure's predictive validity was not supported.

The emergence of a positive correlation with the RMET, while curious, aligns with a body of literature querying the validity of the measure (Peñuelas-Calvo et al., 2019). Notably, findings from confirmatory factor analyses on the RMET are often inconsistent, leading to questions regarding the measure's construct validity (Higgins et al., 2022). For example, the RMET has been postulated to at least partially index other related but distinct constructs, such as emotion recognition (Oakley et al., 2016), fluid intelligence (Navarro Garcia, 2021), and language (Gallant & Good, 2019). However, these issues are not unique to the RMET, with a vast majority of ToM measures also displaying ambiguous psychometric properties (Bosco et al., 2016). There is also contrasting evidence to suggest that the RMET has adequate psychometric properties and is likely indexing some sub-component of ToM (Vellante et al., 2013). Despite this conflicting evidence base, the measure is still widely

adopted (Mintah & Parlow, 2018). Thus, while this positive relationship with the AQ highlights growing concerns regarding the validity of the RMET, it remains one of the only plausible measures of ToM for use with neurotypical adults.

Positive correlations between the RMET, SST-B, and AQ may also be due to limitations associated with the AQ. Recent research has indicated that the AQ may be an unreliable measure of traits associated with autism spectrum disorder when used in non-clinical samples (Jia et al., 2019). Alternatively, some evidence has suggested that autism spectrum disorder may not be universally associated with impaired ToM (Gernsbacher & Yergeau, 2019). Gernsbacher and Yergeau note that some individuals with autism can pass ToM tasks, an impossible feat should ToM deficits be complete and universal. They query whether communication deficits that characterise autism may instead account for poor performance on linguistically complex ToM tasks. As individuals with autism spectrum disorder vary in communication deficits, they assert that individuals with superior communication abilities can pass these ToM tests. The present study's focus on subclinical traits of autism spectrum disorder, which is often associated with relatively minor social deficits (Pisula & Ziegart-Sadowska, 2015), may account for the observed non-significant relationships.

Currently, it is unclear which conclusion is supported. Future research may look to replicate these analyses with an alternative measure of autism spectrum disorder traits. If significant negative relationships emerge between this alternative measure of autism spectrum disorder traits and the RMET and SST-B, this will imply that using the AQ with neurotypical populations was potentially inappropriate. In contrast, if positive or no relationships emerge, this would support the literature suggesting that autism spectrum disorder traits are not universally associated with ToM deficits (Gernsbacher & Yergeau, 2019).

Regardless, the present results do not support the assertion that the SST is indexing ToM, nor can they be taken to suggest that the SST-B and RMET are not indexing ToM. Alternatively, the results imply that the SST is comparatively better than the SST-B and RMET in predicting the presence of autism spectrum disorder symptomology, as measured by the AQ, in this sample of neurotypical adults. While a non-significant negative relationship between the SST and AQ supports the overall validity of the SST, in isolation, it is insufficient to conclude that participant scores on the SST are indexing ToM.

Alternatively, the presence of a sufficiently strong positive relationship between performance on the SST/SST-B and the RMET would provide greater support for the SST or SST-B's likelihood to reflect ToM relative to the predictive validity evaluations. While both the SST and SST-B positively correlated with the RMET, the strength of this relationship was stronger for the SST-B ($r_s = .46$) than the SST ($r_s = .24$). Additionally, the strength of the relationship between the RMET and SST fell below the preregistered acceptable criterion of $r_s \geq .30$. This initially suggests that the SST-B is doing a comparatively better job of reflecting ToM than the SST.

Independent ToM subscale exploratory factor analyses for the SST and SST-B further support the comparative superiority of the SST-B for measuring ToM. Independent analyses of the SST-B ToM subscale indicate that seven of eight items are adequately loading ($\geq .30$) onto factor one. Combined with the concurrent validity analysis, the presence of a single plausible, coherent factor suggests the SST-B as likely indexing some component of ToM. However, these results also indicate that the one item that inadequately loaded (question ten) likely requires refinement or deletion to improve the SST-Bs construct validity.

Although combined ToM subscale exploratory factor analysis suggests common variance between the SST and SST-B, some analyses suggest that these scales are not

uniform in their measurement of the constructs of interest. The SST independent analyses showing the presence of a single plausible, coherent factor (i.e., six of eight items adequately loading onto factor one), suggest some evidence of a unitary factor structure. However, there is less support for the SST as indexing ToM given the strength of the correlation between the SST and the RMET compared to the SST-B scale. Further supporting this conclusion, while factor loadings across independent analyses for the SST and SST-B were of comparable strength, seven items loaded adequately on the SST-B independent analyses versus six on the SST independent analyses. Subsequent factor score analyses also showed that the factor one scores for SST-B had stronger correlations with the RMET than the factor one scores for the SST. Thus, exploratory factor analyses and validity analyses suggest that, within the context of this study, the SST-B may be performing comparatively better in reflecting some aspect of ToM than the original SST.

This finding contrasts the initial validation of the SST, which found a substantially stronger correlation between the SST and RMET (r = .49; Dodell-Feder et al., 2013). While this is plausibly due to the adaption from a verbal to a written format, it does not account for the comparatively stronger correlation between the SST-B and RMET found in the present study. Given that the RMET theoretically measures affective ToM, stronger correlations may be due to a measure requiring greater use of one's affective ToM abilities. However, seven of eight questions on the SST theoretically require affective ToM abilities, while six of eight questions on the SST-B should also employ such abilities. Given that both measures should theoretically measure affective ToM to a similar degree, differences may be due to the SST-B being more difficult, of comparatively greater validity, or the result of other general methodological differences between studies.

Another unexpected finding was the emergence of a significant positive correlation between the SST-B comprehension subscale and scores on the RMET. This indicates the

possibility that, to the extent that the RMET is a valid measure of ToM, the comprehension questions may be indexing some form of ToM ability. If this were true, plausibly, full-scale exploratory factor analyses would be expected to reveal comprehension items on the SST-B clustering with the ToM subscale items, a trend that was observed across combined and independent exploratory factor analyses. When combined, four comprehension items (questions two, three, four, and thirteen) on the SST-B loaded at ≥ .30 on factor one, while only one question on the SST did. Similarly, independent analyses showed three items from the SST-B loaded at ≥ .30 on factor one, while only two items on the SST did.

This finding begs the question: why might these items be indexing ToM? The third SST-B question asks, “Do the adjutant’s actions suggest he is hardworking or lazy? What makes you think that?” Plausibly, this question requires participants to attribute an intention to a behaviour, a common ToM ability (Schaafsma et al., 2015), thereby potentially accounting for the high loading of the question on the ToM factor. Similarly, question four on the SST-B asks, “What does the adjutant mean when he says, “Be soft, Pinin … The major is sleeping?” The marking rubric for this question required the participants to consider the influence one character's actions may have on another character to achieve full marks. Consideration of multiple characters' intentions was only needed to achieve full marks for ToM questions on the SST, not comprehension questions. Thus, the scoring of question four on the SST-B aligning with the scoring of ToM questions may account for its unexpectedly high loading.

Furthermore, question thirteen (“The story is called ‘A Simple Enquiry.’ What is the title referring to?”) also had a loading ≥ .30 on factor one. To correctly answer question thirteen, participants must understand that the Major had made an implicit sexual proposition to his subordinate, an assumption that would potentially require invoking the use of ToM abilities. Additionally, the title “A Simple Enquiry” reflects the use of irony. The ‘enquiry’

was a proposition, and this proposition was not simple. The identification of irony requires utilising ToM abilities (Mitchley et al., 1998). However, it is unclear why question two ("The major has a saucer of oil for what purpose?") had a loading ≥ .30, as the answer to this question is explicitly mentioned in the story. Regardless, these findings suggest that the comprehension questions of the SST-B require refinement or omission due to them indexing the same latent construct as the ToM subscale.

Reliability. To examine the reliability of the SST and SST-B, pre-registered alternate forms, inter-rater, and internal consistency reliability analyses were conducted. Regarding alternate forms reliability, the preregistered acceptable ICC was set at .50. In retrospect, this value is too low to indicate true alternate forms. The minimum acceptable value for tests used for non-clinical judgement is commonly set at .70 (Holmefur et al., 2009; McMaster et al., 2009). The ICC value falling short of this minimum .70 cut-off suggests that these measures were not alternate forms of one another.

The emergence of a suboptimal ICC value between two scales that appear to be indexing different constructs within this study is perhaps somewhat unsurprising. Also likely contributing to suboptimal alternate forms reliability was the process of measure creation. Traditionally, alternate tests are created by constructing a pool of items and splitting this into two or more forms. This was not possible, given that the SST-B required the use of a different story and was constructed at a later date. Instead, a process of item cloning was used. Item cloning involves matching each item on the original measure to an identical item on the new measure (Clause et al., 1998). As both stories examined different emotional states and characters displayed different methods of non-verbal communication, cloned question were unlikely to be entirely equivalent. These factors likely contributed to the SST and SST-B being poor alternate forms.

Inter-rater reliability evaluations were conducted to determine how consistent markers were in applying scores to the SST and SST-B. The process of determining inter-rater reliability aligned with the initial psychometric analyses of the SST conducted by Dodell-Feder et al. (2013), with analyses being conducted at the subscale level for both the SST and SST-B. However, a direct comparison cannot be made to Dodell-Feder et al. as the form of ICC used in their analysis was not reported. Instead, ICC estimates were interpreted in reference to criteria outlined by Koo and Li (2016), which indicated that inter-rater reliability estimates fell between moderate-good (.57 and .75) for the SST and good-excellent (.75 and .92) for the SST-B.

While ICC values for both measures fell short of the anticipated optimal reliability (≥ .80), these findings indicate that for both the ToM and comprehension subscales, the SST-B allowed different raters to apply identical scores more consistently to participants comparative to the SST. Minor revisions to the marking rubric may allow for easier delineation between 0, 1, and 2-point responses resulting in improved inter-rater reliability. Alternatively, providing future raters with systematic training sessions may also result in improvements (Atkinson & Murray, 1987). More extensive changes should also be made to the SST, given the comparatively lower ICC values across both subscales. Without these amendments, the SST-B currently appears to allow for greater consistency in the application of participant scores compared to the SST.

Cronbach's alpha was calculated to provide an estimate of the interrelatedness of items within the context of this study and its sample. Within this study, Cronbach's alpha for both subscales on the SST and SST-B fell below the generally accepted criterion of $\alpha = .70$ (Gliem & Gliem, 2003). However, this was anticipated and reflected in preregistration cut-offs of $\alpha = .25$ and $\alpha = .45$, given that small alpha values were obtained during the initial psychometric analyses of the SST (Dodell-Feder et al., 2013). While alpha tended to fall

close to or above preregistered cut-offs, these pre-registered cut-offs are, in hindsight, relatively arbitrary. Findings still suggest that, within this sample, items across both measures were poorly interrelated and, thereby, it is concluded that both measures had inadequate internal consistency reliability.

Plausibly, violation of tau-equivalence may have contributed to this. Cronbach's alpha assumes that all scale items are equally contributing variance to the latent variable. When this assumption is violated, Cronbach's alpha tends to underestimate internal consistency (Dunn et al., 2014). Exploratory factor analytic findings suggest a violation of this assumption, indicating sample estimates of alpha within this study may be underestimations of the alpha statistic. Violation of these assumptions is common, leading researchers to recommend using McDonald's omega as an alternative to alpha (Dunn et al., 2014). McDonald's omega is less restrictive and allows variances to vary, meaning that omega is often a more accurate estimate of internal consistency. Omega was not adopted in the present study to not further deviate from pre-registered protocol. However, future studies should consider the adoption of omega over alpha to improve accuracy in reliability estimation.

Additional Analyses. Irrespective of other psychometric properties, an important characteristic of the SST and SST-B, if they are to hold utility in the measurement of individual differences in neurotypical adults, is their ability to display an absence of ceiling effects and produce varied scores. Both criteria were met in this study. The range of observed participant scores on the SST and SST-B suggested neither measure displayed ceiling effects, and both had variation in the scores received across the ToM subscales.

Also of interest was the relationship between performance on the comprehension subscales of the SST/SST-B and their respective ToM subscales. Given that poor performance on the comprehension scale was used to justify omission of some individual's

data on the ToM subscale, it was questioned whether poor performance on the comprehension subscale would be associated with poor performance on the ToM subscale. Significant positive correlations emerged for both measures, suggesting that ToM subscale scores were related to comprehension scores. However, given that earlier analyses highlighted that the construct validity of the SST ToM subscale and SST-B comprehension subscale was questionable, the ability to make further interpretations regarding the nature of the relationship between these two variables is limited.

Examining the Relationship between ToM and Video Game Engagement in Relation to the GAM and GLM

While engagement with different video game genres and social contexts was not related to performance on the SST ToM subscale, these findings must be interpreted in relation to the psychometric analysis of the SST within this study. Specifically, it was concluded that within this study, the SST did not appear to be a valid measure of ToM or the purported ToM subprocesses the SST was designed to measure. Considering this, the absence of a relationship between these variables should not be used to conclude that video game engagement and ToM abilities are unrelated.

Alternatively, there is evidence to suggest that within this study, the SST-B may be slightly more valid in its ability to index ToM. Similarly, some wider literature suggests that the RMET may reflect components of ToM (Vellante et al., 2013). Thereby, rerunning earlier analyses using these measures allows for tentative conclusions to be drawn regarding relationships between ToM abilities and video game engagement. However, these analyses were not preregistered and should therefore be interpreted with an appropriate degree of caution.

Findings indicated that playing different video game genres was not related to ToM abilities measured by the SST-B. These results align with those of Kühn et al. (2019), who found that playing video games that included violence was not associated with changes in ToM abilities. Alternatively, these results partially contrast those of Bormann and Greitmeyer (2015), who suggested that playing an adventure game with narration led to improvements in ToM as measured by the RMET. In the present study, a significant positive correlation did not emerge between the RMET and adventure video game engagement in this study. This finding suggests that an alternative explanation may be warranted.

Instead, specific gameplay elements (narration) may cause short-term changes in ToM ability while superordinate differences (genre) do not. Additionally, Bormann and Greitymeyer measured affective ToM ability immediately after 20 minutes of gameplay, whereas the present study did not measure time since the last period of gameplay, indicating that perhaps affective ToM may only improve short-term, and these differences are not long-lasting enough to withstand more lengthy breaks between playing and ToM measurement. Another possibility is that differences in the measures of ToM employed in the two studies may be responsible for the divergent findings. Future studies may benefit from investigating the short-term influence of specific video game gameplay features on ToM instead of focusing on genre or long-term effects.

Interestingly, while a significant positive correlation did not emerge between the RMET and adventure video game engagement, significant negative correlations emerged between the RMET and engagement with fighting and action video games. This finding aligned with the SST-B analyses when gameplay time was controlled, which also observed a significant weak negative correlation with fighting game engagement. However, in the case of the SST-B findings, this reflects a change in r_s = -.18 when not controlling playtime to r_s = -.19 when controlling for playtime. Thus, the effect of dosage appears to be trivial at best.

Regardless, these findings provide some support for the GAM in predicting changes in ToM. Specifically, the GAM would predict that prolonged exposure to violence in fighting and action video game games would lead to increased accessibility of aggressive thoughts and subsequent reductions in trait empathy and emotional responses to these violent scenes (Anderson et al., 2010; Anderson & Bushman, 2018). As empathy and ToM commonly work in tandem (Cerniglia et al., 2019), decreases in empathy would plausibly be associated with concurrent reductions in affective ToM, thereby accounting for the observed negative correlations. However, several other genres that commonly include violent video game play (e.g., virtual worlds, MMOG's, survival horror, strategy, adventure, action, role play) did not observe significant negative correlations with the RMET or SST-B. Perhaps fighting and action games may involve a higher degree of violence resulting in the observed correlations, or perhaps the findings are Type I error due to the high number of analyses. As such, replications are required to support the GAM in predicting changes in ToM as a result of video game play.

Alternatively, analyses utilising the SST-B when gameplay time was not controlled do not appear to support the GAM or GLM. Contrasting the GAM, significant negative correlations also did not emerge between SST-B ToM scores and engagement with many genres that may include violence. Alternatively, the GLM would predict a range of improvements and decrements in ToM across different genres dependent upon the cognitive processes that are repeatedly employed. As this was also not observed, neither theory appears to be supported by the current study's results. These contrasting findings across measures and analyses highlight the current gap in the video game literature regarding appropriate theory for making predictions about changes, or lack thereof, in cognitive abilities, particularly ToM abilities. Development, refinement and articulation of media psychology theories that are

broader than the features of violence in video games are therefore required to overcome this gap.

It must also be noted that 56 correlations were calculated during these preregistered and exploratory analyses. Subsequently, a number of false positives are anticipated (Simas et al., 2014). These findings may represent Type I error since only three significant correlations were observed, all during the exploratory phase of the analyses. Thus, all of the findings, particularly the exploratory findings, warrant subsequent replication before any firmer conclusions are drawn.

Regarding the relationship between video game social context and ToM, analyses utilising the RMET and SST-B suggested that playing either multi-player, single-player, or both video game types was unrelated to ToM abilities. While no wider literature has investigated this effect, the GLM would predict that repeated learning encounters through discourse and interaction in multi-player video games should likely lead to improvements in ToM over time. Additionally, developmental literature suggests that greater ToM competence is associated with higher social engagement (Fujita & Itakura, 2008). Thus, it was anticipated that higher multi-player video game engagement would be associated with greater ToM abilities. The absence of this relationship further highlights the inability of existing theories to adequately account for observed findings and signals the need for more nuanced theories within this area.

Summary and Future Directions

The primary findings of Study 1 are summarised below in reference to the overarching aims for Study 1 outlined in Chapter 4:

1. *To develop and pilot an alternate version of the Short Story Task with neurotypical adults.*

While the SST-B was not an adequate alternate form of the SST, the SST-B may be a slightly more valid measure of ToM relative to the SST. However, adjustments should be made to scoring training procedures and the marking rubrics of the SST and SST-B to improve consistency in score application. While internal consistency reliability was inadequate in this sample, this may partially be the result of using Cronbach's alpha over McDonald's omega. Thus, future studies should adopt omega.

2. *To determine whether engagement with different video game genres is associated with differences in the ToM abilities of neurotypical adults.*

    2.1. To determine whether the social context of video game play is related to differences in an individual's ToM abilities.

Some evidence indicated that engagement with fighting and action video games might be related to poorer performance on ToM measures. However, these analyses were exploratory and potentially reflected Type I error. Further, social context was not related to ToM abilities.

3. *To determine whether the GLM or GAM proposes a more accurate theoretical framework for investigating the influence of video game play on ToM abilities.*

    3.1. To determine whether relationships between ToM ability and the social context of video games are consistent with predictions made by the GLM.

Overarchingly, existing media psychology theory appeared to be insufficient for investigating relationships between ToM and video game play. However, some evidence suggested that predictions made by the GAM were accurate in some areas (e.g., fighting and action video game engagement may be related to performance on some ToM measures). However, predictions made by the GLM were not consistent with analyses examining the relationship between the social context of video game play and ToM.

These findings highlight several avenues for future research. Regarding the SST and SST-B, future research is warranted to replicate the present findings, given that preliminary evidence suggests the SST-B may be slightly more valid than the SST in the measurement of ToM. Notably, the present study would benefit from some specific improvements to methodology – in particular, expanding upon exploratory factor analysis findings using confirmatory factor analysis to explore the construct validity of these measures further; using an alternative measure of autism spectrum disorder symptomology given concerns about the validity of using of the AQ with neurotypical populations; adjustment of rater training procedures and/or measure marking rubrics to improve score application consistency; and adoption of McDonald's omega over Cronbach's alpha as a measure of internal reliability.

Similarly, the video game analyses also require replication due to their exploratory nature and the risk that Type I error may have contributed to the observed findings. However, these analyses require critical consideration of the ToM measure used, given that this study also highlighted that existing ToM measures often have questionable psychometric properties (Bosco et al., 2016). Further, the video game literature would also benefit from articulation and testing of new theoretical frameworks, given that these results highlight limitations in the utility of the existing theories.

CHAPTER 8:

STUDY 2 INTRODUCTION

Originally, it was planned that Study 2 would expand upon the findings of Study 1 by examining the following aims:

1. To causally determine whether engagement with a specific video game genre (identified in Study 1) is associated with differences in the ToM abilities of neurotypical adults.
2. To determine whether the GLM or GAM proposes a more accurate theoretical framework for investigating causal changes in ToM abilities as a result of video game play.

However, the results of Study 1 indicated that investigating these originally proposed aims may not have been a fruitful avenue forward as:

a) The Short Story Task-B (SST-B) was not an adequate alternate form of the Short Story Task (SST).
b) Exploratory video game analyses only resulted in three significant correlations between Theory of Mind (ToM) measures and video game genre engagement which may reflect Type I error given the number of analyses run.

Regarding a), this would mean a post-test only design would need to be adopted as opposed to the initially intended pretest post-test design. While this would still allow for conclusions to be drawn regarding the potential presence of causal relationships, the strength of these conclusions would be much weaker than initially hoped. In addition, b) brought into question the likelihood of an experimental study producing significant findings. Together, these findings cast doubt on the utility of proceeding with an experimental paradigm for Study 2.

Further, data collection for Study 2's experiment was set to commence during the second quarter of 2020. At the end of the first quarter of 2020, New Zealand entered a nationwide Level 4 lockdown in relation to the Covid-19 pandemic (Summers et al., 2020). Ongoing restrictions on movement and social gathering imposed by the New Zealand government and Massey University meant that collecting experimental data during 2020 was unlikely to be feasible. These covid restrictions, in tandem with findings from Study 1, thereby necessitated a change in the scope and aims of Study 2. As such, the scope, aims, method, data analysis procedures, and preregistration for Study 2 were rapidly adjusted under lockdown restrictions. To foreshadow, this contributed to the preregistration of aims and analyses, which, in hindsight, were overly ambitious and/or less viable than initially expected. Limitations and deviations from the preregistered protocol due to this process are explicitly outlined throughout subsequent chapters.

While findings from Study 1 suggested that further exploration of the relationship between video games and ToM was unlikely to be fruitful, preregistered and exploratory analyses during Study 1 highlighted that the SST-B may be an incrementally better measure of ToM relative to the SST. Further investigation of the psychometric properties of these measures was therefore warranted. Thereby, the aims and scope of Study 2 were readjusted, and the primary aim of Study 2 became:

1. To determine whether the Short Story Task or the alternate form piloted during Study 1 is a better measure of ToM with neurotypical adults.

To address Aim 1, and align with Study 1, Study 2 adopted an online survey design and replicated the initial preregistered psychometric analyses of the SST and SST-B. To expand upon Study 1, Study 2 also looked to utilise confirmatory factor analysis to examine the SST and SST-B's construct validity (DiStefano & Hess, 2005). Further, findings from

Study 1 also brought into question the utility of using the Autism Spectrum Quotient (AQ) with neurotypical adults given both the RMET and SST-B positively correlated with the AQ. As such, an alternative measure of autism spectrum disorder symptomology was to be adopted during Study 2 to determine whether this finding generalised to another measure, or whether this finding was unique to the AQ. As such, Study 2 looked to use the Rivto Autism and Asperger's Diagnostic Scale (RAADS-14) as opposed to the AQ in investigating the predictive validity of the SST and SST-B.

The results of Study 1 also raised questions about the utility of the the comprehension subscale included in the SST and SST-B. This is compounded by a growing desire for a reduction in the administration time of psychometric tests, in both research and practise, as a cost-reduction method (Yates & Taub, 2003). Thereby, it was speculated that perhaps the comprehension subscales could be replaced with a simplified self-report measure of story understanding to reduce the SST and SST-B's administration time.

Similarly, there is a growing desire to substitute tests of cognitive ability (e.g., ToM tests) for short self-report questionnaires (Yates & Taub, 2003). Again, the primary motivation for this change is to reduce researcher/clinician input and administration time to further cost reductions. Additionally, there is evidence to suggest that metacognitive judgements about one's ability can prove accurate in some areas of cognition such as memory strength (Weber & Brewer, 2004). Although, in many domains, especially with regard to social interactions, confidence can be a relatively poor predictor of performance and often exhibits overconfidence (e.g., Dunning et al., 1990). While metacognitive judgements about individuals' ability have been examined across a wide range of domains (see Lichenstein et al., (1982) for a review), to date, little work has examined the degree to which individuals have metacognitive knowledge about their own ToM skills. Although some recent research has suggested that individuals have good insight about their ToM abilities using self-report

measures (Crehan et al., 2020; Hutchins et al., 2021), the majority of these measures were developed subsequent to data collection and remain relatively lengthy (i.e., 48-60 items). Thus, I also aimed to assess whether a shorter single-item measure might be adequate to capture individuals' ToM. Thereby, Aim 1.1. of Study 2 was:

> 1.1. To determine whether self-report scales of ToM ability and short story comprehension are adequate substitutes for tests of ToM ability and short story comprehension.

While research examining links between ToM ability and video games is scant and mixed (Bormann & Greitemeyer, 2015; Kühn et al., 2019), an area of growing interest is in links between reading literary fiction and improvements in ToM. It was initially posited that engagement with fictional storytelling provides readers access to content about human social interaction and psychological processes that they would not otherwise have in their daily lives (Mar & Oatley, 2008). Access to this information was thought to improve an individual's social knowledge, thereby improving their ability to make accurate attributions about the mental states of others. In addition to improving an individual's social knowledge, it has also been theorised that reading literary fiction improves the processes that may underpin ToM (Oatley, 2016). Reading literary fiction is thought to involve mental state simulation of story characters, thereby providing readers with extra practice in mental state attribution. This assertion is supported by neuroimaging research, as Mar (2011) observed overlap in the neural regions activated during fiction reading and ToM engagement, suggesting fiction reading may invoke ToM abilities. However, alternative explanations may be warranted. For example, individuals with relatively superior ToM abilities may enjoy reading more literary fiction.

Irrespective of how or whether literary fiction engagement may theoretically improve ToM, correlational evidence suggests that individuals who are more familiar with literary fiction perform better on measures of ToM (Mumper & Gerrig, 2017). Subsequent experimental research further supported the causal relationship between reading literary fiction and short-term improvements in ToM (Dodell-Feder & Tamir, 2018). However, despite these promising initial findings, failed experimental replications have been conducted (Panero et al., 2016; Samur et al., 2018).

Panero et al.'s (2016) and Samur et al.'s (2018) contrasting findings highlight a growing problem within the area of psychological research – the replication crisis (Pashler & Wagenmakers, 2012; Schooler, 2014). In short, the replication crisis refers to a phenomenon in which a large proportion of significant findings, within the psychological and wider scientific literature, cannot be reproduced (Ioannidis, 2005). The reasons for the replication crisis are many and varied, with methods of data collection, analysis, and reporting, all being implicated (Simmons et al., 2016). The replication crisis warrants researchers replicating findings across time and contexts to verify their accuracy. Further, to ensure that data analysis strategy is not responsible for false positive results, the preregistration of data collection and analysis plans are recommended (Simmons et al., 2016). Given that replications investigating causal links between literary fiction and ToM have produced mixed findings (Panero et al., 2016; Samur et al., 2018), the second and final aim investigated in this pre-registered study was to:

2. Replicate the finding that familiarity with literary fiction is associated with greater ToM abilities (Mumper & Gerrig, 2017).

CHAPTER 9:

STUDY 2 METHOD

This chapter will initially outline the edits made to the SST-B resulting from Study 1. Following this, measure selection, data collection procedures, and preregistered statistical analyses will be summarised.

SST-B Edits. Minor edits to the marking rubric of the SST-B were made during Study 1. These changes increased the number of participant responses that would be assigned a one-point score for questions five, six, and seven, which are observable in Table 9.1. Edits were made as it was observed during Study 1 that some participants correctly identified mental states that were not accounted for in the original rubric. As such, these participants were originally awarded a zero. As such, the resulting edits were made, meaning these responses were now allocated a score of one. These edits were not used while scoring the Study 1 SST-B results.

Table 9.1

Original and Edited Marking Rubric Criteria for Assigning One-Point to Questions Five, Six, and Seven on the SST-B

Original Criteria	New Criteria
5. Pinin is trying to deflect the major's question (no reason as to why he may be doing this is given)	5. Pinin is trying to deflect the major's question (no reason as to why he may be doing this is given) or incorrect reasoning given; because the questions make him uncomfortable – no mention of deflection.
6. Makes mention of the action intention (i.e., to make Pinin talk more openly/protect the majors/Pinin's privacy) without acknowledging emotions (e.g., put them at ease/make Pinin less nervous etc).	6. Makes mention of the action intention (i.e., to make Pinin talk more openly/protect the majors/Pinin's privacy/as he is going to proposition him) without acknowledging emotions (e.g., put them at ease/make Pinin less nervous etc) OR just mentions emotions without intentions.
7. Any answer that does not accurately interpret the major's perception of Pinin's intention; failure to acknowledge that it is the major's perception of Pinin's intention (e.g., Pinin believed that homosexuality was inferior).	7. Any answer that only partially interprets the major's perception of Pinin's intention (e.g., He believes Pinin looks down on him); failure to acknowledge that it is the major's perception of Pinin's intention (e.g., Pinin believed that homosexuality was inferior; Pinin believed he was better than the Major).

Participants. This study is a replication and extension of Study 1. In line with this, an identical sample size of 112 participants was adopted in the present study. The original sample size was calculated using G*Power. An a priori power analysis was set at .90 power to detect a medium effect size of r = .30 at the .05 alpha error probability, which indicated that a sample size of 112 participants was required. An r = .30 was selected in line with Cohen's (1988) effect size conventions to detect a 'medium' effect size.

One hundred and twelve participants ranged in age from 18 to 65, with the mean age being 33 (± 11.4). One individual failed to specify their age. Males (56%) accounted for the majority of participants, while females (42%), those who preferred not to say (1%), and Other

(Please Specify; 1%) accounted for the remaining. No specification was given for the individual who selected Other (Please Specify). Individuals came from a range of educational backgrounds, with completion of a bachelor's degree (44%) being the most common. Further demographic information is observable in Table 9.2.

Table 9.2

Demographic Characteristics of Participants in Study 2

Characteristic	Frequency (N=112)	Percent (%)
Age (Years), Mean (SD)	33.3 ± 11.4	
Gender		
Male	63	56.3
Female	47	42.0
Prefer not to say	1	0.9
Other (Please Specify)	1	0.9
Education		
Early Childhood	0	0
Primary	1	0.9
Lower Secondary	1	0.9
Upper Secondary	20	17.9
Post-Secondary (Non-Tertiary)	5	4.5
Short Cycle Tertiary	15	13.4
Bachelor's Degree	49	43.8
Master's Degree	20	17.9
Doctoral Degree	1	0.9

Note. SD = Standard Deviation.

Inclusion and Exclusion Criteria. Inclusion and exclusion criteria and associated rationales for these are the same as for Study 1, with two revisions. Firstly, the video game

playtime criterion was removed as the present study did not seek to replicate the video game findings of Study 1. Second, participants were excluded if they participated in Study 1.

Thus, to be eligible to take part in the present study, participants were required to:

1. Have not participated in Study 1.
2. Be currently living in New Zealand, Australia, Canada, the United States, or the United Kingdom.
3. Complete the survey on a desktop computer.
4. Be over the age of 16.
5. Not have a severe visual impairment.
6. Have no current or previous diagnoses of neurological, developmental, or psychological disorders.

Participant Recruitment. All participants were recruited online using Prolific. Prolific is an online survey-hosting website where potential participants are informed about studies they may self-select to partake in. Outlined inclusion criteria were implemented using Prolific's custom pre-screening tool. Participants were only informed about the study if they completed studies on a desktop and were from New Zealand, Australia, Canada, the United States, or the United Kingdom. Participants were redirected to an information sheet (Appendix C-2) if they expressed interest. If they consented to participate, they were prompted to click to the next page of the survey.

Measures. In addition to the SST-B, a range of measures were selected to examine the outlined research objectives. These measures and the rationale for their selection will be discussed. Rationales for the RMET and SST, measures used in Study 1, can be read on pages 70 – 72 of the present thesis. In the present study, internal consistency reliability for the RMET was $\omega = .79$.

Author Recognition Test. The Author Recognition Test (ART) was initially developed to measure an individual's familiarity with fictional print media (Stanovich & West, 1989). The measure presents participants with the names of 65 real authors and 65 fake authors. Participants must identify the names of familiar authors. Participants are encouraged to only select authors they know, as one point is given for each correct identification, and one is removed for each incorrect identification. An updated version of the ART was used in the present study, given that author familiarity changes over time (Acheson et al., 2008). However, Acheson et al.'s version of the ART only reported the real authors used and not the distractors. Distractors used in the present study were gleaned from Martin-Chang and Gould (2008). As Martin-Chang and Gould provided 75 distractors while only 65 were needed, ten were randomly omitted. For a list of real authors, retained distractors, and omitted distractors, see Appendix D.

Scores on the ART highly correlate with reading behaviours (R. F. West et al., 1993) and show greater validity than commonly used self-report measures (Mol & Bus, 2011). The measure also displays good levels of split-half reliability (r = .86; Ocal et al., (2017)). Notably, the version of the ART adopted in the present study has been used in literature examining correlations between ToM abilities and literary fiction familiarity (Kidd & Castano, 2013). In the present study, internal consistency reliability for the ART was α = .94.

Ritvo Autism & Asperger Diagnostic Scale. The Ritvo Autism & Asperger Diagnostic Scale (RAADS-14) is a short form of the Ritvo Autism and Asperger Diagnostic Scale-Revised. The RAADS-14 presents participants with 14 self-report items designed to screen for traits associated with the autistic spectrum (Eriksson et al., 2013). Participants are given four response options to these items: true now and when I was young; true only now; true only when I was younger than 16; and never true. The measure has three subscales: mentalizing deficits, social anxiety, and sensory reactivity. An endorsement of 'true now and

when I was young' receives a score of three, 'true only now' receives two, 'true only when I was younger than 16' receives one, and 'never true' receives zero except for on item six which is reverse coded. This allows for a total score of 42.

Eriksson et al.'s (2013) pilot study supported the internal consistency of the RAADS-14 (α = .90) and suggested it could discriminate between ASD and non-psychiatric populations (area under curve = .99). A more recent study, published subsequent to the conduction of Study 2, supported the convergent (r = .81 with the Autism Spectrum Quotient – 10) and the discriminant validity (r = .75 with the EQ-Short) of the RAADS-14 but brought into question its construct validity, specificity, and reliability (Kember & Williams, 2021). While conclusions drawn by Kember and Williams highlight limitations in the validity of the RAADS-14 within New Zealand, Prolific only hosts a small number of users from New Zealand (approximately 0.003%). Thereby, it is unlikely that a significant proportion of the sample in Study 2 were from New Zealand, meaning that findings by Kember and Williams were unlikely to pose a significant limitation in adopting the RAADS-14. Thus, the RAADS-14 presented an adequate alternative to the AQ in evaluating the predictive validity of the SST and SST-B in the present study. In the present study, internal consistency reliability for the RAADS-14 was ω = .82.

Procedure. After consenting to participation, participants were prompted to enter their Prolific ID to match response sets to associated Prolific submissions. This was to ensure that only participants who completed the study received monetary compensation. Participants were counterbalanced to complete either the SST or SST-B first. All participants completed both the SST and SST-B. After reading the instructions and story, participants were asked, "Have you read this story before?" If 'Yes' was selected, participants would then be asked: "How long ago did you read it?", "What do you remember about the story?" and "Have you discussed the story with anyone?". If 'No' was selected, these questions did not appear. One

participant indicated they had read *The End of Something* before. Follow-up answers indicated they were referring to reading the story during the study. Participants were then asked to rate on a one to seven scale, “How well do you think you understood the story you just read?”. The subsequent 13 comprehension and ToM questions were presented on the same page as the story so participants could refer back to it if necessary. An additional attention check question was added after item six on the SST and after item 13 on the SST-B (“I read instructions carefully. To show that you are reading these instructions, please leave this question blank.”). If this question was incorrectly responded to, listwise deletion was used for the participant’s data.

All participants were then presented with the ART. This was to reduce potential practise effects that completion of the SST/SST-B would have on the later presented counterbalanced measure by maximizing time between measure presentations. This allowed for the maximum amount of memory interference and decay between the two measures. Keeping this questionnaire in the same position standardized this timeframe across participants. All authors’ names were presented in alphabetical order. Participants were instructed to click on a name if they knew this was a real author.

The presentation of the RMET was the same as in Study 1. The RMET was presented in the third position for all participants. It was hypothesised that completing the SST/SST-B may improve performance on the RMET due to the effects of reading literary fiction on ToM (Dodell-Feder & Tamir, 2018). By standardising the presentation, this effect was consistent across all participants. The RMET was prefaced with a task description and a list of mental state descriptors that would be shown across the task. Participants could then select those they did not know and receive an associated definition. Upon task completion, participants were asked on a ten-point scale how difficult they found the task, how much mental effort it

required, and how well they thought they did. They were also asked to indicate if they had completed this task before.

Participants were then presented with the SST or SST-B, depending on which measure was completed earlier in the study. Participants were then asked to rate on a one to seven scale “How well do you think you understand other people’s beliefs, intentions, and emotions?”. Following this, participants completed the RAADS-14. The RAADS-14 was presented last as it has a high degree of face validity regarding its measurement of traits of autism spectrum disorder. Some literature indicates that autism spectrum disorder may be associated with deficits in ToM abilities (Baron-Cohen, Wheelwright, Hill, et al., 2001). Therefore, to mitigate demand effects on participants’ ToM abilities, it was completed after all ToM measures. A third attention check was imbedded after question seven on the RAADS-14 (“I have once owned a three headed dog.”) If participants responded with any response other than “Never True”, listwise deletion was used for the participant’s data.

Participants were then asked their age, gender, level of education, and whether they were proficient English speakers. All participants indicated they were proficient English speakers. Following this, participants were asked whether they encountered any issues with the survey and to guess what the study hypotheses may be. This was to check whether potential awareness of study aims had impacted or biased responses. Finally, participants were provided with a debrief section explaining the study's aims and whether they would like to be informed of the study’s results later.

Preregistered Analysis Strategy. All statistical analyses were preregistered on the Open Science Framework. Confirmatory Factor Analyses and calculations of McDonald’s Omega were conducted in JASP version 0.16.1. The remaining analyses were conducted in SPSS version 25.0. Data were initially screened for missing data using Little’s (1988)

Missing Completely At Random (MCAR) test. As missing data was MCAR, the expectation-maximization algorithm was used. Data was also inspected for normality and the presence of outliers. Outliers were defined as ± 3.29 standard deviations from the mean (Tabachnick et al., 2007). Skewness and kurtosis were calculated for all measures to determine the appropriateness of parametric tests, assess normality, and determine the presence of ceiling effects. Spearman and Pearson's correlations were calculated between the SST and SST-B ToM and respective comprehension subscales to determine whether story comprehension was related to performance on the ToM subscale

Statistical analyses for the psychometrics properties of the SST and SST-B followed those outlined in Study 1. These were primarily based on procedures outlined Dodell-Feder et al. (2013). For interrater reliability analyses, all forms were marked by the primary researcher. Then, 25% of the forms were randomly selected to be marked by another independent marker. The second marker was a Doctor of Philosophy candidate in Food Technology at the Massey University School of Food and Advanced Technology and was provided with a brief training session on scoring the SST and SST-B. Intra-class Correlation Coefficients (ICC) estimates, based upon an absolute agreement model, were then calculated independently for the comprehension and ToM subscales. These were interpreted in reference to guidelines outlined by Koo and Li (2016). Estimates were not compared to values obtained by Dodell-Feder et al. for the SST, as the form of ICC used was not reported.

Internal consistency reliability was measured using McDonald's omega. Omega was adopted over Cronbach's alpha as omega does not assume all items have equal factor loadings (tau-equivalence). As tau-equivalence is not commonly observed, Omega often reflects a more accurate estimate of reliability (Hayes & Coutts, 2020). While Cronbach's alpha is not a perfect analogue for omega, omega was preregistered to be considered acceptable if $\omega \geq .45$ for the ToM subscales and $\omega \geq .25$ in line with values of alpha observed

by Dodell-Feder et al. (2013) during original psychometric analyses of the SST. However, traditional interpretation guidelines for omega would suggest that $\omega \geq .70$ to be considered acceptable (Hussey & Hughes, 2020).

Mirroring Study 1, concurrent validity for the SST and SST-B was measured using Spearman's Rank Order Correlation's between the SST and SST-B ToM subscales and the RMET. Spearman correlations were used in place of Pearson correlations as the RMET data was significantly skewed. The standard alpha value of .05 was used to determine whether the SST and SST-B ToM subscales displayed concurrent validity with the RMET if $p < .05$. Dodell-Feder et al. (2013) found a medium effect size of $r = .49$ between the RMET and SST. Additionally, Study 1 observed an $r = .46$ between the RMET and the SST-B. As such, concurrent validity was determined to be acceptable if $r \geq .30$. Spearman correlations were also calculated between the SST and SST-B comprehension subscales and the RMET to determine whether the comprehension questions indexed mental state abilities.

Predictive validity for SST and SST-B was measured by calculating Pearson correlations between the RAADS-14 and the SST and SST-B ToM subscales. These values were then squared to determine how much variability in RAADS-14 scores was accounted for by performance on these measures. While the RAADS-14 is not a diagnostic tool, a score of 14 or above on the RAADS-14 may indicate the presence of autism spectrum disorder, which in some individuals may be associated with impairment in ToM abilities (Eriksson et al., 2013). Given this, an analysis with and without individuals who scored above this threshold was conducted to analyse whether these individuals scores affected the predictive validity of the SST and SST-B. The standard alpha value of .05 was used to determine whether the SST and SST-B ToM subscale predicted scores on the RAADS-14. Additionally, given that correlations between the SST, SST-B, and AQ in Study 1 were low, effect sizes were anticipated to be small. Given this, predictive validity was deemed acceptable if scores

on the SST or SST-B ToM subscale accounted for ≥ 1% of the variation in RAADS-14 scores.

Preregistered confirmatory factor analysis hypotheses looked to compare the fit of two models for the SST and SST-B against one another. One model fixed all item loadings to 1, while the other used loadings informed by previous exploratory analyses. These models are observable on this study's Open Science Framework preregistration (https://doi.org/10.17605/OSF.IO/WTEPN). Subsequently, it was determined that this was an ineffective use of confirmatory factor analysis and would not aid in further understanding of the construct validity of these measures. Instead, confirmatory factor analysis was run using the diagonally weighted least squares estimation method to examine the fit of independent SST and SST-B models. These models were informed by preregistered models, whereby all ToM subscale items loaded onto one factor for each measure. Factor variance was fixed to 1 for these models. Unstandardised parameter estimates are reported.

The following non-preregistered fit indices were used to evaluate model fit: Comparative Fit Index (CFI); Root Mean Square Error of Approximation (RMSEA); Tucker Lewis Index (TLI); and Standardised Root Mean Square of the Residual (SRMR). These indices were interpreted in relation to conventional criteria outlined by Hu and Bentler (1999). However, Hu and Bentler's criteria were developed for continuous data using the maximum likelihood estimation method. Current evidence suggests that alternative criteria are likely required, as the diagonally weighted least squares estimation method tends to overestimate CFI and TLI while underestimating RMSEA (Xia & Yang, 2019). However, to the author's knowledge, no such alternative criteria currently exist.

Additionally, while absolute guidelines suggest that $n = 100$ is sufficient for the conduction of confirmatory factor analysis (Kline, 2015), there is contrasting evidence to

suggest that when $n < 200$, overestimation may be further pronounced (DiStefano & Morgan, 2014). This is a contentious area, with researchers arguing that overestimation may be negligible even when $n < 200$ (Savalei & Rhemtulla, 2013), while others state that fit indices are still overestimated even when $n = 500$ (Xia & Yang, 2019). Regardless, a conservative approach would suggest that using the diagonally weighted least squares estimation method alongside smaller sample sizes will increase the likelihood of concluding adequate model fit despite this not being the case.

Bayesian Information Criterion and Chi-squared statistics were originally preregistered for use to compare between model fit. Given that alternative models were not computed, these fit indices could not be interpreted in this manner. While the Chi-squared statistic can theoretically be used to interpret model fit, it has several limitations, including its sensitivity to sample size and a restrictive standard of perfect fit (Kline, 2015). As such, its use for the interpretation of model fit is not commonly recommended, with other fit indices mentioned above holding greater utility (Hussey & Hughes, 2020). Thus, while Chi-squared was calculated and reported in line with preregistered protocol, it was not used to interpret model fit.

To determine whether participants could accurately predict their ToM abilities, Spearman correlations were calculated between self-reported ToM abilities and the RMET, SST ToM subscale, and the SST-B ToM subscale. The standard alpha value of .05 was used to determine whether these three measures correlated with scores on the self-report ToM questions if $p < .05$. RMET analyses were not preregistered. As there was no literature to guide anticipated effect sizes, correlations of $r_s \geq .10$ were preregistered to be considered of interest.

Similarly, to determine whether participants could accurately report their understanding of the stories used in the SST and SST-B, Spearman correlations were calculated between self-reported story comprehension and actual performance on the respective comprehension scale of the SST or SST-B. Ultimately, this was to ascertain whether the SST and SST-B's administration time could be reduced by removing the comprehension questions if participants were accurate reporters of their story understanding. The standard alpha value of .05 was used to determine whether scores on these two subscales correlated with scores on the self-report story comprehension if $p < .05$. As there was no literature to guide anticipated effect sizes, correlations of $r_s \geq .10$ were preregistered to be considered of interest.

Spearman correlations between the ART, the RMET, the SST ToM subscale, and the SST-B ToM subscale were calculated to ascertain whether familiarity with literary fiction was associated with improved performance on measures of ToM ability. The standard alpha value of .05 was used to determine whether scores on these three measures correlated with scores on the ART if $p < .05$. Meta-analytic results suggest that reading literary fiction is associated with small positive improvements in performance on tests of ToM ability (Mumper & Gerrig, 2017). In accordance with Cohen's (1988) conventions for small effect sizes, $r_s \geq .10$ was employed as a cut-off.

Factor scores were then calculated for the preregistered informed SST and SST-B models. Outlined validity analyses were then reconducted using the generated factor scores. Correlations between the RAADS-14 and factor scores did not use the data set with imputed values.

CHAPTER 10:

STUDY 2 RESULTS

Data Screening

Subsequent statistical procedures and interpretive guidelines were preregistered on the Open Science Framework (https://doi.org/10.17605/OSF.IO/WTEPN). Deviations from this preregistration and exploratory analyses are explicitly outlined.

Statistical Outliers. Total and appropriate subscale scores for all measures were analysed for univariate outliers. In line with current convention, outliers were deemed scores that fell outside the 3.29 Standard Deviations of the mean (SD; Leys et al., 2019; Tabachnick et al., 2007). The SST and SST-B comprehension subscales plausibly function as a control task by assessing story understanding. Individuals are anticipated to perform at or near ceiling. Pairwise deletion was used if individuals scored below four points on the SST on the comprehension subscales, as a score below four was arbitrarily deemed to indicate inadequate story engagement. Finally, listwise deletion was used for participants' data if they incorrectly responded to an awareness check question.

No outliers were detected on the Author Recognition Test (ART) and Ritvo Autism & Asperger Diagnostic Scale (RAADS-14). One outlier was detected on the Reading the Mind in the Eyes Test – Revised Edition (RMET). Pairwise deletion was used for this data point. For the SST, seven participants scored below four on the comprehension subscale; for the SST-B, four participants scored below four. Pairwise deletion was used for these data points.

Eight participants failed the attention check embedded within the RAADS-14. Their data was deleted and not included in subsequent analyses. Additionally, one participant completed the survey twice. Their second submission was deleted, while their initial submission was retained.

Missing Data. Every item was analysed for missing values. One data point was missing for the SST question "How well do you think you understood the story you just read?" and 14 total data points were missing across questions one, three, five, eight, nine, eleven, twelve, and thirteen on the RAADS-14. Little's (1998) Missing Completely at Random test indicated that these data points were most likely missing completely at random. Given that these missing data points represented less than 5% of the overall item scores, the expectation-maximization algorithm was used to impute data in SPSS version 25 to improve statistical power (Scheffer, 2002).

Of the 103 participants' data used for subsequent analyses, 17 (16.5%) failed to answer any items on the RMET. As with Study 1, participant comments indicated this was likely due to survey design or incompatibility with some participants' computers. As all items for the RMET were missing for these individuals, they were excluded from analyses involving the RMET.

Skewness, Kurtosis, and Test Selection. The application of parametric tests requires data to follow a normal distribution. Therefore, the symmetry and pointedness of the data were calculated using z-scores based on skewness and kurtosis (Table 10.1). In line with conventions outlined by (Kim, 2013), a Z-score of 3.29 or greater was used to evidence a departure from normality (as n = 50-300). Resultantly, non-parametric tests were used for analyses involving the SST comprehension subscale, the RMET, and the ART.

In line with Study 1's results, scores across both measures showed substantial variation. The absence of total scores greater than 12 across both measures also suggested an absence of ceiling effects (Figure 10.1).

Table 10.1

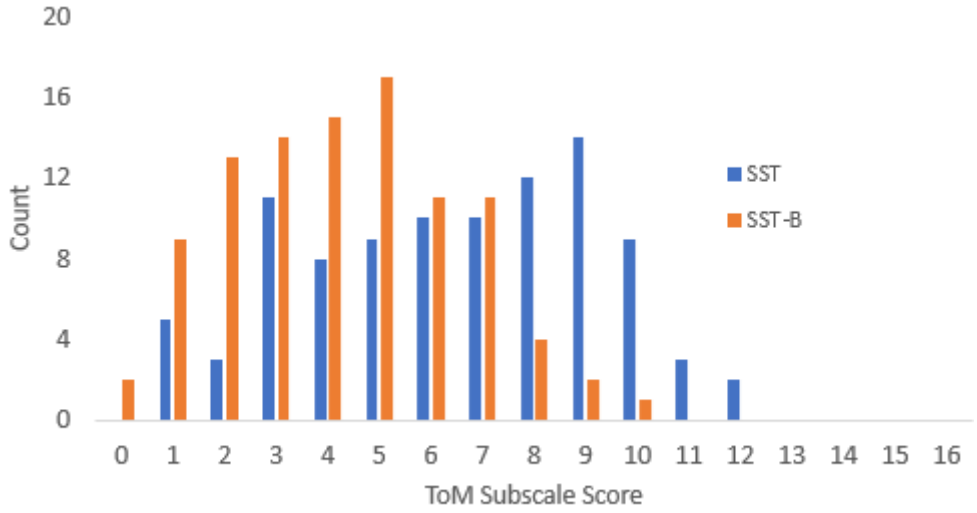
Skewness and Kurtosis Indices for Measures and Subscales

Measures	n	Skewness			Kurtosis		
		Statistic	z-score	p	Statistic	z-score	p
SST							
Comprehension	96	-1.10	-4.48	<.001	0.68	1.40	.16
Theory of Mind	96	-0.18	-0.73	.47	-0.91	-1.86	.06
SST-B							
Comprehension	99	-0.12	-0.50	.61	-0.86	-1.79	.07
Theory of Mind	99	0.20	0.83	.41	-0.56	-1.16	.25
RMET	86	-1.31	-5.03	<.001	2.07	4.04	<.001
ART	103	0.96	4.02	<.001	0.57	1.21	.23
RAADS-14	103	0.33	1.38	.17	-0.48	1.02	.31

Note. Standard error of the mean for skewness and kurtosis were .25 and .49 for the SST; .24 and .48 for the SST-B, respectively; .26 and .51 for the RMET, respectively; and .24 and .47 for the ART and RAADS-14.

Figure 10.1

Distribution of the Theory of Mind Subscale Scores for the SST and SST-B



Practice Effects. Before conducting analyses involving the RMET, a Mann Whitney U test was conducted to determine whether self-reported familiarity with the RMET was associated with improvements in test performance. As can be seen from Table 10.2, test familiarity was not associated with improved performance on the RMET.

Table 10.2

Mann-Whitney U Results Comparing Previous Completion of the RMET and RMET Total Score

Familiarity	n	Mean Rank	U	p
Familiar	10	41.3	358.0	.77
Unfamiliar	76	43.8		

Story Comprehension and ToM Subscale Performance. Performance on the comprehension subscales of the SST and SST-B were correlated against their respective ToM subscales. This was to determine whether story comprehension subscale performance was related to performance on the ToM subscale. Spearman and Pearson correlation coefficients are observable in Table 10.3. Across both measures, significant positive correlations of a moderate strength emerged between ToM and their respective comprehension subscales across both the SST and SST-B, suggesting that improved performance on one subscale is related to improved performance on the other.

Table 10.3

Spearman and Pearson Correlations Between the SST and SST-B Comprehension Subscales and their Respective ToM Subscale

Measure	Correlation (95% CI)	p
SST	.43 (.25, .58)	<.001
SST-B	.38 (.19, .54)	<.001

SST-B correlations were calculated using Pearson correlations. SST correlations were calculated using Spearman correlations.

Reliability Evaluation

Inter-Rater Reliability. All inter-rater reliability estimates were based upon a single measure, absolute agreement, two-way random effects model. Intraclass Correlation Coefficients (ICC) and associated 95% confidence intervals are observable in Table 10.4. Across both subscales, the SST displayed greater ICC estimates than the SST-B. For the SST and SST-B, comprehension subscale scores fell in the 'excellent' range, while ToM subscale scores fell in the 'moderate' range (Koo & Li, 2016). It should be noted that while SST-B ToM ICC estimates fell in the 'good' range for Study 1, this only reflects an absolute difference in strength of .04 between studies.

Overall, these findings are curious as ICC estimates for the SST-B were weaker than those observed during Study 1. Conversely, SST ICC estimates for Study 2 were stronger. In the present study, the marker was given a brief training period, whereby they were taught how to score both the SST and SST-B. This training period was not provided in Study 1 and may account for the more consistent scoring of the SST during this study. Similar effects may not have been observed for the SST-B as ICC estimates may have already been at or near their ceiling, given that the measure was designed for written administration/scoring.

Table 10.4

Inter-Rater Reliability Estimates for the Total Scale Scores of the SST and SST-B Comprehension and ToM Subscales

Measure	ICC (95% CI)	p
SST		
Comprehension	.98 (.94 , .99)	<.001
Theory of Mind	.72 (.14 , .90)	<.001
SST-B		
Comprehension	.91 (.81 , .96)	<.001
Theory of Mind	.71 (.01 , .91)	<.001

Note. CI = Confidence Interval.

Internal Consistency Reliability. Internal consistency for the SST and SST-B was calculated using McDonald's Omega. Dodell-Feder et al. (2013) observed Cronbach's alpha values of $\alpha = .54$ (ToM subscale) and $\alpha = .31$ (comprehension subscale) on the SST, which they attributed to the range of content and varying question difficulties within these subscales (Dodell-Feder et al., 2013). Omega was thereby anticipated to be low, with preregistered acceptable values being $\omega \geq .45$ for the ToM subscales and $\omega \geq .25$ for the comprehension subscales. However, it is acknowledged that, in hindsight, these values are too low to reflect adequate internal consistency. Thus, values will also be interpreted in reference to common conventional criteria, whereby ω will be deemed acceptable at ≥ 0.70 (Hussey & Hughes, 2020; Nunnally, 1994). While caution was applied in interpretation, given that this was not preregistered, the adoption of conventional criteria poses a more conservative cut-off.

Table 10.5 presents the omega coefficients for the SST and SST-B. Both subscales for both measures fell above the preregistered acceptable criteria. However, only omega for the SST subscales was acceptable based on conventional criteria (i.e., $\omega \geq 0.70$). These findings provide some support for the internal consistency of the SST subscales but not the SST-B.

Table 10.5

McDonald's Omega Total for the SST and SST-B

Measure	ω (95% CI)
SST	
Comprehension	.70 (.52 , .80)
Theory of Mind	.71 (.63 , .78)
SST-B	
Comprehension	.57 (.32 , .68)
Theory of Mind	.55 (.33 , .66)

Note. CI = Confidence Interval. Number of items in ToM subscale = 8; Number of items in the comprehension subscale = 5.

Validity Evaluation

Concurrent Validity. Spearman correlations between the RMET and the SST and SST-B ToM subscale were calculated to determine the concurrent validity of the SST and SST-B. To determine whether performance on the RMET was related to story comprehension, Spearman correlations were calculated between the SST-B comprehension subscale scores and the RMET. Concurrent validity was deemed acceptable for the SST and SST-B if $r_s \geq .30$.

Spearman correlations are observable in Table 10.6. In line with Study 1 and preregistered criteria, the SST-B ToM subscale showed a significant correlation with the

RMET and had an $r_s \geq .30$. However, the SST ToM subscale did not significantly correlate with the RMET and did not exceed $r_s \geq .30$. While the failure to exceed $r_s \geq .30$ aligns with Study 1, failure to significantly correlate with RMET stands in contrast to the findings of Study 1. However, this only reflects a difference of $\Delta r_s = .05$ between studies. The SST-B comprehension subscale significantly correlated with the RMET, and this aligns with the results of Study 1 and suggests that the SST-B comprehension subscale may inadvertently be indexing some aspects of ToM. However, the significant correlation between the SST comprehension subscale and the RMET contrasts Study 1's findings and also suggest the SST comprehension subscale may inadvertently be indexing some aspects of ToM.

Table 10.6

Spearman Correlations Between the SST and SST-B Subscales and the RMET

Measure	Correlation (95% CI)	p
SST		
Comprehension	.48 (.30 , .63)	<.001
Theory of Mind	.19 (-.03 , .39)	.09
SST-B		
Comprehension	.24 (.03 , .43)	.03
Theory of Mind	.32 (.11 , .50)	.003

Note. CI = Confidence Interval. Two-tailed.

Predictive Validity. Pearson correlations between the RAADS-14 and the SST and SST-B ToM subscale were calculated and squared to determine the predictive validity of the SST and SST-B. Spearman correlations were calculated between the RMET and RAADS-14 as a reference. Given the results observed in Study 1, predictive validity was anticipated to be

low. Thereby, if scores on the SST or SST-B ToM subscale accounted for ≥ 1% of the variation in RAADS-14 scores, predictive validity was deemed to be acceptable.

Results are observable in Table 10.7. With the inclusion of individuals scores above the clinical threshold, the SST and RMET showed a trend in the predicted direction with an acceptable R^2 = .01 and .06. This suggests that participant scores on the SST and RMET respectively account for 1% and 6% of the variation in participants' scores on the RAADS-14. However, contrasting the hypothesised outcome, the SST-B showed a non-significant positive correlation with the RAADS-14. The observed results for the SST and SST-B align with those observed in Study 1, whereby the SST met preregistered acceptable criteria while the SST-B did not. However, the RMET did not have a significant negative correlation with the Autism Spectrum Quotient (AQ) in Study 1. Thus, the significant negative correlation between the RAADS-14 and the RMET in this study contrasts the results observed in Study 1. The adoption of the RAADS-14, as opposed to the AQ, may partially account for the observed difference.

Analyses were also conducted after removing individuals' data who scored above the clinical cut-off (≥14) on the RAADS-14. This was not preregistered and resulted in a remaining *n* = 44 for the RAADS-14. This observation aligns with findings by Kember and Williams (2021), which suggested that the RAADS-14 may have poor specificity. A significant negative correlation emerged between the SST-B and RAADS-14 with an R^2 = .13. Both correlations for SST and RMET reduced in size, with R^2 = .00 and .03, respectively. However, given the limited sample sizes used for these analyses, findings should be interpreted cautiously.

Table 10.7

Pearson and Spearman Correlations between the SST ToM Subscale, SST-B ToM Subscale, RMET and the RAADS-14

Measure	Clinical Threshold Included				Clinical Threshold Excluded			
	n	Correlation (95% CI)	R^2	p	n	Correlation (95% CI)	R^2	p
SST	96	-.09 (-.29 , .11)	.01	.38	42	-.03 (-.33 , .28)	.00	.86
SST-B	99	.05 (-.15 , .24)	.00	.66	43	-.36 (-.60 , -.07)	.13	.02
RMET	86	-.25 (-.44 , -.04)	.06	.02	36	-.16 (-.46 , .18)	.03	.35

Note. CI = Confidence Interval. The RAADS-14 threshold for clinical significance is ≤ 14. SST and SST-B correlations were calculated using Pearson correlations. RMET correlations were calculated using Spearman correlations.

Confirmatory Factor Analyses. Independent confirmatory factor analyses were conducted for the SST and SST-B using models informed by preregistered models (i.e., all ToM subscale items loading onto one factor). Non-preregistered fit indices (CFI, RMSEA, TLI, and SRMR) were calculated to aid in the determination of model fit. Fit measures were interpreted in line with commonly accepted criteria outlined by Hu and Bentler (1999; CFI > 0.95; RMSEA < 0.06; TLI > 0.95; SRMR < 0.09). However, it is acknowledged that these criteria likely need to be more conservative given this study's sample size and the method of estimation adopted (Xia & Yang, 2019). Additionally, given the highlighted deviations from preregistered protocol, caution in finding interpretation is further warranted. Confirmatory factor analysis fit indices for the SST and SST-B, based upon preregistered models, are observable in Table 10.8. Measurement models for both the SST and SST are observable as Figures 10.2 and 10.3.

No fit indices suggested adequate model fit for the SST-B model, while all fit indices, excluding SRMR for the SST model, fell within the acceptable range. A model's fit should only be considered *good* if all inference criteria fall within an acceptable range (Hussey & Hughes, 2020). As such, neither model is plausible. These findings may be due to the retention of items in these models that displayed factor loadings <.30 during exploratory factor analysis in Study 1.

Table 10.8

Confirmatory Factor Analyses Fit Measures for the Full SST and SST-B ToM Subscales

Measure	CFI	RMSEA	TLI	SRMR	BIC	χ^2	df	p
SST	.99	.03	.99	.10	-	21.5	20	.37
SST-B	.78	.12	.70	.16	-	50.56	20	<.001

Note. BIC = Bayesian Information Criterion

Figure 10.2

Measurement Model Including Unstandardised Estimates for the SST Based Upon Preregistered

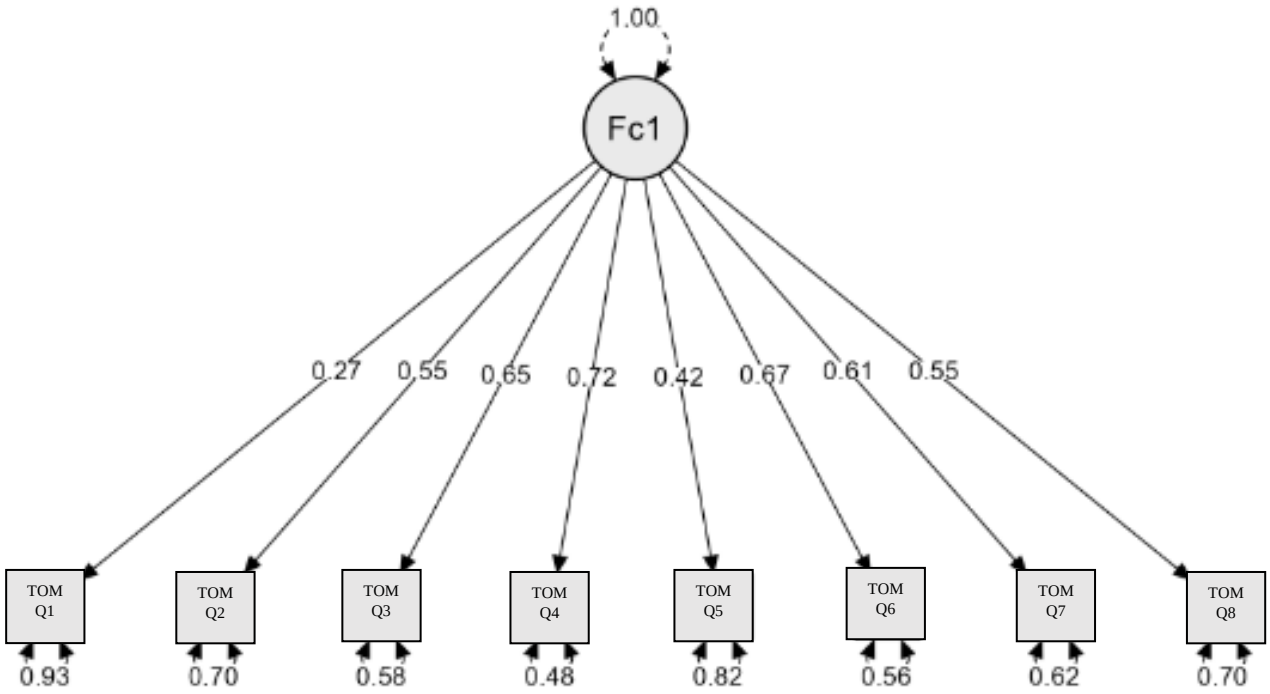
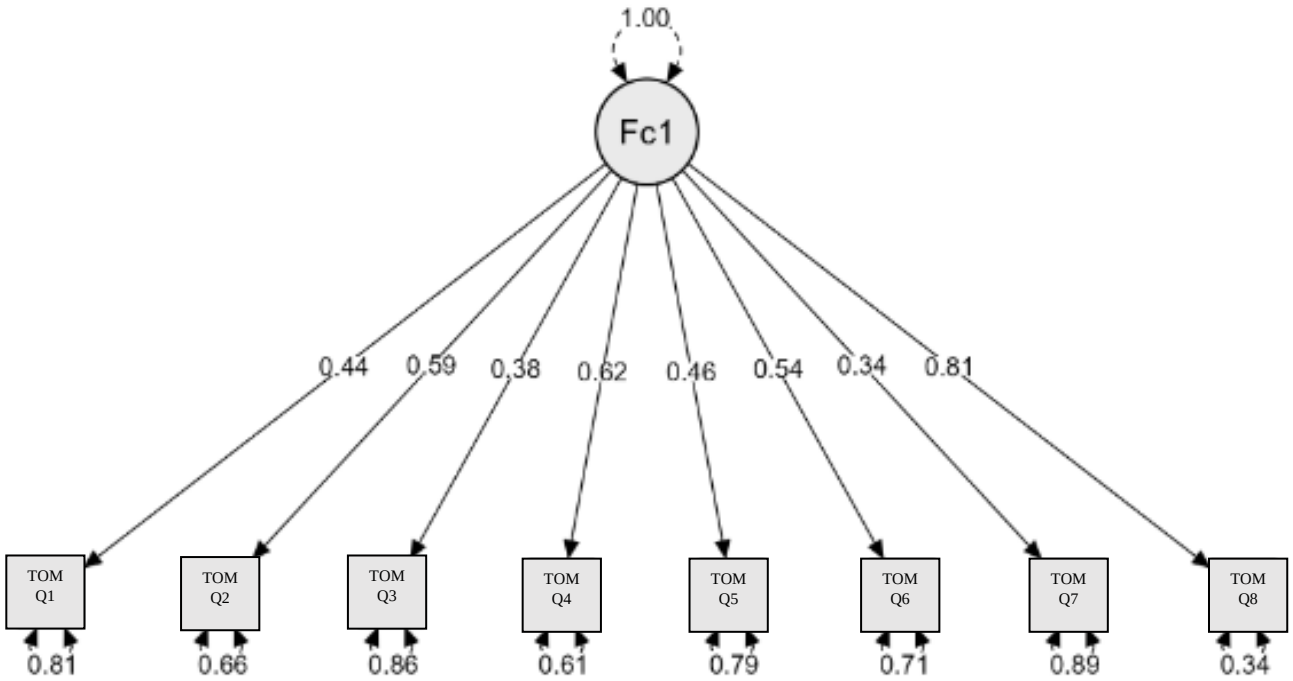


Figure 10.3

Measurement Model Including Unstandardised Estimates for the SST-B Based Upon Preregistered Models



Self-Reported ToM and Story Comprehension Analyses

Self-Reported ToM Ability. Spearman correlations were calculated between self-reported ToM abilities and the RMET, SST ToM subscale, and the SST-B ToM subscale. As there was no literature to guide anticipated effect sizes, correlations of $r_s \geq .10$ were preregistered as being of interest as the traditional cut-off for a small effect (Cohen, 1988).

Results are observable in Table 10.9. These findings suggest that individuals' self-reported perception of their own ToM abilities does not correlate with performance on the SST, SST-B, and RMET. While the correlation between self-reports and scores on the RMET was ≥ .10, this correlation was non-significant.

Table 10.9

Spearman Correlations Between Self-Reported ToM Abilities and Performance on the SST ToM Subscale, SST-B ToM Subscale, and RMET

Measure	Correlation (95% CI)	p
SST	.05 (-.15 , .25)	.61
SST-B	-.07 (-.27 , .13)	.47
RMET	.15 (-.06 , .35)	.16

Note. CI = Confidence Interval.

Self-Reported Story Comprehension. Spearman correlations between the self-reported story understanding and the SST SST-B comprehension subscales. Again, as there was no literature to guide anticipated effect sizes, correlations of $r_s \geq .10$ were preregistered as being of interest.

Results are observable in Table 10.10. Participants self-reported understanding of *The End of Something* significantly correlated with their performance on the comprehension subscale for the SST. Alternatively, participants' self-reported understanding of *A Simple Enquiry* did not correlate with their performance on the comprehension subscale of the SST-B.

Table 10.10

Spearman Correlations Between Self-Reported Story Comprehension and Performance on the SST and SST-B Comprehension Subscales

Measure	Correlation (95% CI)	p
SST	.20 (.00 , .39)	.04
SST-B	-.01 (-.21 , .19)	.90

Note. CI = Confidence Interval.

Literary Fiction Familiarity and ToM Abilities.

Spearman correlations were calculated between the ART, the RMET, the SST ToM subscale, and the SST-B ToM subscale. Small effect sizes were anticipated and as such, $r_s \geq .10$ were preregistered as being of interest.

Results are observable in Table 10.11. All correlations were significant, ranging in strength from $r_s = .32 - .41$. Interestingly, the observed effect sizes were stronger than anticipated. However, the anticipated effect sizes, based upon findings by Mumper and Gerrig (2017), did fall within the confidence intervals for both the RMET and SST.

Table 10.11

Spearman Correlations Between Scores on the SST ToM Subscale, SST-B ToM Subscale, RMET and the ART

Measure	Correlation (95% CI)	p
SST	.32 (.12 , .49)	.001
SST-B	.41 (.23 , .56)	<.001
RMET	.32 (.11 , .50)	.003

Note. CI = Confidence Interval.

Factor Score Analyses.

It was preregistered that factor score would be calculated based upon the model which showed the best fit, and validity analyses rerun, to further explore the construct validity of these measures. However, given preregistered competing models were not computed, factor scores were calculated from the SST and SST-B full-scale ToM single-factor models based on these preregistered models. Interpretation of findings is limited by these deviations from preregistered protocol, in addition to the observation that neither model showed adequate fit.

Results examining the concurrent validity of the SST and SST-B utilising factor scores are observable in Table 10.12. In line with earlier results (Table 10.6), the SST-B ToM subscale showed a significant correlation with the RMET and had an $r_s \geq .30$. Alternatively, the correlation between the SST factor scores and the RMET remained below preregistered acceptable criteria ($r_s \geq .30$) and was non-significant.

Table 10.12

Spearman Correlations Between the SST and SST-B ToM Subscale Factor Scores and the RMET

Measure	Correlation (95% CI)	p
SST	.18 (.-.03 , .38)	.09
SST-B	.34 (.14 , .52)	.002

Note. CI = Confidence Interval. Two-tailed.

Results examining the predictive validity of the SST and SST-B utilising factor scores are observable in Table 10.13. Aligning with earlier results (Table 10.7), the SST negatively correlated with performance on the RAADS-14 while the SST-B positively correlated. Both correlations remained non-significant. Again, in line with earlier analyses, the omission of RAADS-14 scores above the clinical cut-off (≥14) resulted in a significant negative correlation emerging for the SST-B while the strength of the correlation for the SST reduced.

Table 10.13

Pearson Correlations between the SST and SST-B ToM Subscale Factor Scores and the RAADS-14

Measure	Clinical Threshold Included				Clinical Threshold Excluded			
	n	Correlation (95% CI)	R^2	p	n	Correlation (95% CI)	R^2	p
SST	96	-.11 (-.30 , .10)	.02	.18	42	-.03 (-.33, .27)	.00	.83
SST-B	99	.07 (-.13 , .26)	.00	.47	43	-.33 (-.58 , -.04)	.11	.03

Note. CI = Confidence Interval. The RAADS-14 threshold for clinical significance is ≤ 14.

Exploratory Analyses

Confirmatory Factor Analyses of Non-Preregistered Models. Confirmatory factor analysis requires the pre-specification of empirically or theoretically informed models. Models which omitted items that displayed factor loadings <.30 during Study 1 were not prespecified. Despite this, post hoc modifications are routinely made and are acceptable if certain criteria are satisfied (Bowen & Guo, 2011). Namely, post hoc changes are justified if prespecified models show inadequate fit and there is either empirical or theoretical evidence to justify the changes (Bowen & Guo, 2011). In this case, both criteria are met. Thus, items five and twelve on the SST and item ten on the SST-B were omitted, and analyses were rerun. Results are observable in Table 10.14. Measurement models for the SST and SST-B are observable in Figures 10.4 and 10.5.

For the SST, all fit indices indicated good fit. While superficially, it could be concluded that this indicates the model is plausible, it must be acknowledged that 1) these analyses and models were not preregistered and require replication, 2) these values are likely overestimated by the method of estimation, and 3) the sample size utilised likely further

contributed to overestimation. Findings should therefore be cautiously interpreted. While the SST-B findings are beholden to these same limitations, SRMR fell below acceptable criteria meaning model fit remained inadequate (Hussey & Hughes, 2020).

Table 10.14

Confirmatory Factor Analyses Fit Measures for the SST and SST-B ToM Subscales Omitting Items with Factor Loadings <.3

Measure	CFI	RMSEA	TLI	SRMR	BIC	χ^2	df	p
SST	1.0	.00	1.0	.08	-	8.2	9	.52
SST-B	.98	.03	.97	.12	-	16.0	14	.34

Note. BIC = Bayesian Information Criterion. For the SST question 5 and 12 were omitted; for the SST-B question 10 was omitted.

Figure 10.4

Non-preregistered measurement model including unstandardised estimates for the SST

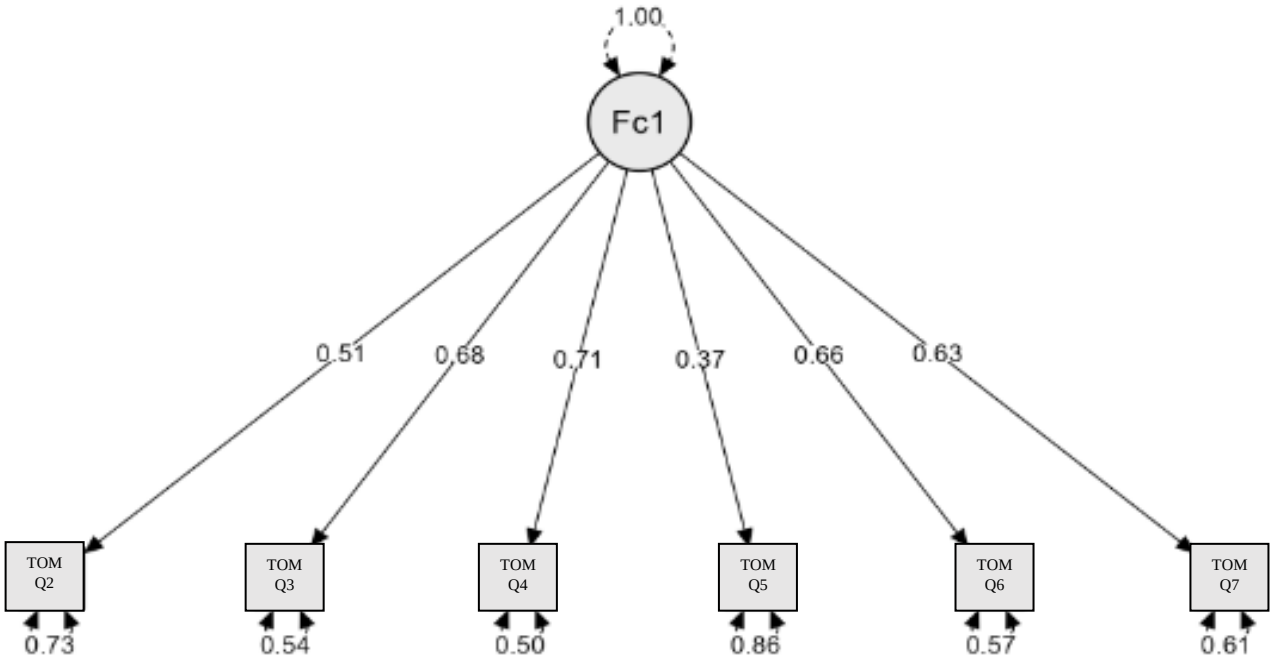
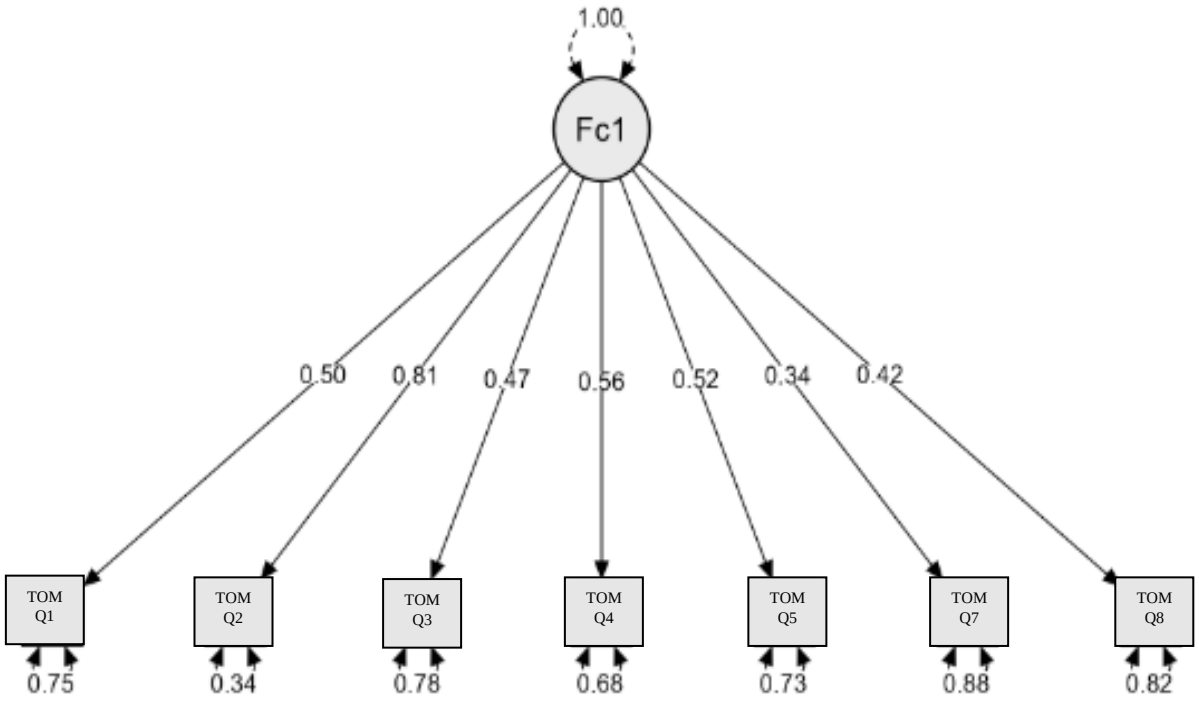


Figure 10.5

Non-Preregistered Measurement Model Including Unstandardised Estimates for the SST-B



CHAPTER 11:

STUDY 2 DISCUSSION

The rationale for Study 2 was largely based upon the results of Study 1 in the present thesis, which indicated that the Short Story Task-B (SST-B) was not an adequate alternate form of the Short Story Task (SST). However, preliminary evidence from Study 1 also suggested that the SST-B may be an improved measure of Theory of Mind (ToM) relative to the SST. Thus, the primary aim of Study 2 was to replicate the psychometric analyses from Study 1 and extend findings using confirmatory factor analysis. The SST and SST-B both contain a comprehension subscale to determine whether the reader has adequately engaged with the story's content. Another aim of this study was to determine whether an individual's self-reported story comprehension was an adequate predictor of performance on the comprehension subscales to reduce administration time in future applications of the SST and SST-B.

Furthermore, it was investigated whether an individual's self-reported ToM abilities adequately predicted performance on purported tests of ToM ability. This was to determine whether individual had metacognitive knowledge about their own ToM skills, and whether administration time could be further reduced in future research by substituting ability tests for brief self-report measures. Finally, a small but compelling body of research suggests that familiarity with literary fiction is associated with small improvements in ToM abilities (Mumper & Gerrig, 2017). The second aim of this study looked to replicate these findings.

In summary, only the predictive validity of the SST was supported, while only the concurrent validity of the SST-B was supported. However, confirmatory factor analyses indicated that both measures required refinement to improve their construct validity. These findings aligned with reliability analyses, indicating a need for further measure and marking rubric refinement. Regarding Aim 1.1 of this study, results indicated that single-item self-

report scales were not adequate substitutes for ToM measures or the SST and SST-B comprehension subscales. Alternatively, significant positive correlations emerged between performance on purported ToM measures and familiarity with literary fiction, replicating Mumper and Gerrig's (2017) findings. These findings will be elaborated on in the following sections.

Examination of the Psychometric Properties of the SST and SST-B

Validity. Validity analyses in Study 2 primarily mirrored those in Study 1 while building upon identified limitations and other findings. During Study 1's predictive validity analyses, it was postulated that the measurement of autism spectrum disorder symptomology in neurotypical adults using the Autism Spectrum Quotient (AQ) may have been inappropriate (Jia et al., 2019). Specifically, it was thought possible that the use of the AQ contributed to the unexpected observed associations between the Reading the Mind in the Test – Revised Edition (RMET) and SST-B, both of which were found to positively correlate with a measure of autism symptomology when a negative correlation was anticipated. Thus, the Rivto Autism and Asperger's Diagnostic Scale (RAADS-14) was adopted in Study 2 to examine whether this unexpected finding was due to the use of the AQ or whether it replicated with a different measure of autism spectrum disorder symptomatology.

Mirroring Study 1's findings, analyses partially supported the predictive validity of the SST as participant scores on the SST ToM subscale accounted for approximately 1% of the variation in participants' scores on the RAADS-14. Contrastingly, the SST-B positively correlated with the RAADS-14, suggesting that the results of Study 2 did not support the predictive validity of the SST-B. This finding suggests that the SST shows greater utility than the SST-B in its ability to tap autism spectrum disorder symptomology as measured by the RAADS-14. However, this conclusion must be balanced against emerging evidence highlighting limitations in the validity of the RAADS-14 with some western populations

(Kember & Williams, 2021). As such, findings regarding the SST and SST-B would benefit from replication once the RAADS-14 is refined or more robust screening tools for autism spectrum disorder symptomology are developed. Overall, given that the correlation between the RAADS-14 and the SST was weak and non-significant, this finding only provides some weak support at best to the validity of the SST.

Also contrasting Study 1's findings, the RMET displayed a significant negative correlation with the RAADS-14, with scores on the RMET accounting for 6% of the variation in participants' RAADS-14 responses. Given the absence of a significant negative correlation between the RMET and AQ in Study 1, this finding supports literature indicating that applying the AQ with neurotypical participants may be inappropriate (Jia et al., 2019).

Curiously, analyses using only participants' scores on the RAADS-14 below the clinical threshold resulted in a significant negative correlation between the RAADS-14 and the SST-B. This finding, while interesting, should not be extrapolated as supporting the predictive validity of the SST-B, as these analyses were not preregistered, and the sample size employed was relatively small (n = 44). However, future research may look to establish whether existing ToM measures may hold differential utility in assessing autism symptomology across different populations and clinical and subclinical levels of symptom severity.

Overall, while the present study's findings further support the predictive validity of the SST relative to the SST-B, in isolation, they are again insufficient to conclude that scores on the SST are indexing ToM, nor do they suggest the SST-B is not. Alternatively, findings from predictive validity analyses across Studies 1 and 2 suggest that the SST may hold greater utility than the SST-B in tapping autism spectrum disorder symptomology in neurotypical adults. Thereby, where researchers or clinicians require a measure for this

purpose, the SST would be more appropriate than the SST-B, though neither measure is ideal for this goal.

Building upon these analyses, concurrent and construct validity evaluations further explored the likelihood that the SST and SST-B reflected ToM. Aligning with findings from Study 1, the SST-B showed a significant positive correlation with the RMET, which fell above the preregistered acceptable criterion of $r_s \geq .30$. While the SST positively correlated with the RMET, the strength of this correlation also fell below preregistered acceptable criteria. This correlation was also non-significant, contrasting findings from Study 1 and the initial psychometric analysis of the SST (Dodell-Feder & Tamir, 2018). However, this only reflects a difference of $\Delta r_s = .05$ between Study 1 and 2 and may be attributable to normal variability between two independent samples. Regardless, the observed consistent, albeit weak, significant positive correlations between the RMET and SST-B across Studies 1 and 2 suggest that the SST-B appears to be performing a comparatively better job of reflecting some aspect(s) of ToM than the SST.

In contrast, findings from confirmatory factor analyses, based upon preregistered models (i.e., all ToM subscale items for each measure loading onto one factor), cast some doubt on the construct validity of both the SST and SST-B. All fit indices for the SST-B indicated an inadequate model fit. While some fit indices for the SST suggested adequate model fit, the Standardized Root Mean Squared Residual (SRMR) did not meet conventional criteria for determining acceptable fit, suggesting that the model is implausible (Hussey & Hughes, 2020). These findings occurred despite model fit indices likely being overestimates of true model fit given the small sample size ($n < 200$) and method of estimation employed (DiStefano & Morgan, 2014; Xia & Yang, 2019). This suggests that when the SST and SST-B ToM subscales are delivered in their originally intended form, they appear to be

inadequately indexing their underlying latent construct. Further interpretation of factor score analyses based upon these models is therefore limited.

Given these findings, confirmatory factor analyses were rerun on models with items exhibiting factor loadings <.30 in Study 1 being omitted. Following this, all fit indices indicated adequate model fit for the SST. However, caution is warranted in the interpretation of these findings. Namely, these models and analyses were not preregistered, with adjustments being made post hoc; fit indices are likely overestimated due to the small sample size and method of estimation (DiStefano & Morgan, 2014; Xia & Yang, 2019); and concurrent validity evaluations across Study 1 and 2 cast doubt on the utility of the SST as a true index of ToM. However, regarding the latter point, it is acknowledged that these concurrent validity analyses were conducted using the full SST ToM scale rather than the refined scale. Regardless, these analyses suggest that following the omission of items five and twelve, there appears to be increased statistical robustness and coherence in what the SST is indexing. The precise nature of the construct being tapped by the refined measure is unclear based on available information. Thereby, further research, replication, and scale development should be conducted, with greater statistical power, to examine whether this factor structure replicates in a new sample and, if so, explore the precise nature of the construct the SST is indexing.

While the SST-B model is beholden to similar limitations outlined above for the SST model (i.e., this model was not preregistered, and the small sample size and estimation method likely contributed to fit overestimation), SRMR still fell below acceptable criterion following item deletion. Taking a conservative approach to evaluating model fit suggests that this adjusted model is inadequate (Hussey & Hughes, 2020). This finding signals a need for further measure and model refinement before the SST-B is utilised as a measure of ToM in future research.

Overall, findings across Studies 1 and 2 suggest limitations in the SST and SST-B's ability to reflect ToM, with findings from confirmatory factor analyses generally highlighting a need for further measure refinement and model testing. This conclusion fits with wider literature, which notes that the majority of commonly employed ToM measures likely have questionable validity and require further refinement (Bosco et al., 2016; François & Rossetti, 2020). This pattern of ToM measures showing questionable psychometric properties, which this thesis further contributes to highlighting, is plausibly related to the conceptual ambiguity of ToM (François & Rossetti, 2020). While this thesis initially proposed an operationalizable framework for ToM, it was acknowledged during the literature review that there is presently no universally accepted definition, theory, or method of operationalising ToM. Without a unifying framework underpinning how ToM is defined and measured, this pattern of ambiguous and perplexing findings is likely to perpetuate (François & Rossetti, 2020).

It is worth acknowledging however that findings from predictive validity analyses were consistent across Studies 1 and 2 and suggest that the SST may hold some utility in measuring autism spectrum disorder symptomology with neurotypical adults. Findings from concurrent validity analyses were also consistent across studies, suggesting that the SST-B may be incrementally better than the SST in the measurement of ToM. However, the contrasting findings across validity analyses indicate that the SST and SST-B should be further refined as scales and their psychometric properties reexplored before they are used in future research to examine ToM. To foreshadow, the General Discussion will further consider reasons why the SST and SST-B have questionable validity and how they may be amended to plausibly improve their validity.

Reliability. To examine different aspects of the reliability of the SST and SST-B, pre-registered inter-rater and internal consistency reliability analyses were replicated. However, building upon identified limitations with the use of Cronbach's alpha in Study 1, McDonald's

omega was adopted as an index of internal consistency. The alternate form reliability evaluations were not replicated due to findings from Study 1 strongly suggesting that the SST and SST-B were not adequate alternate forms.

Interrater reliability analyses were conducted to determine how consistent markers were in applying scores to the SST and SST-B. Analyses mirrored those in Study 1 and the initial psychometric analyses of the SST conducted by Dodell-Feder et al. (2013), with analyses conducted at the subscale level for both the SST and SST-B. ICC estimates were interpreted in reference to criteria outlined by Koo and Li (2016), who indicated that inter-rater reliability estimates for both the SST (.72 and .98) and SST-B (.71 and .91) subscales fell between moderate-excellent.

Observed ICC estimates for the SST-B were generally consistent with Study 1. This finding suggests that amendments made to the SST-B's marking rubric and the inclusion of a rater training period did not substantially improve consistency in score application for the SST-B. Alternatively, estimates for the SST are notably stronger and suggest that the inclusion of this short training period for the marker may have allowed for more consistent scoring on the SST (Atkinson & Murray, 1987). Plausibly, ICC estimates for the SST-B may not have improved due to them already being at or near their ceiling, given that the measure was designed for written administration/scoring while the SST was not.

Despite these changes, reliability for both the SST and SST-B ToM subscales remains sub-optimal (ICC < .80) when interpreted in line with preregistered criteria for Study 1. This suggests that for future written administration of these measures, more significant revisions to the marking rubrics may be required. However, providing future raters with systematic training sessions is recommended, given the stronger observed inter-rater reliability estimates for the SST following the inclusion of a training session. Currently, marking rubrics and rater

training procedures allow for approximately an equal degree of consistency in applying participant scores across the SST and SST-B.

In this study, there was some support for the internal consistency of both SST subscales as $\omega \geq 0.70$, while for both SST-B subscales $\omega < .70$, suggesting internal consistency was inadequate. Given that Study 1 and the initial psychometric analyses of the SST adopted Cronbach's alpha as an estimate of internal consistency (Dodell-Feder et al., 2013), the comparison of findings across studies is limited. However, it requires consideration that for the SST, estimates of omega were close to conventional cut-off criteria, with the lower bound of their confidence intervals falling below .70. This suggests that while the internal consistency of the SST was relatively superior to the SST-B in this study, both measures require improvement. Thus, these findings generally align with earlier analyses, which suggested a need for further measure refinement.

Additional Analyses. Mirroring findings from Study 1, scores on the SST-B comprehension subscale had a significant positive correlation with scores on the RMET. It was proposed during Study 1 that several items on the SST-B comprehension subscale may inadvertently be indexing ToM, which was supported by exploratory factor analyses. As changes were not made to these items between studies, the emergence of this significant positive correlation was anticipated. Contrasting Study 1, the SST comprehension subscale displayed a significant positive correlation with the RMET. This represents a difference in strength of $\Delta r_s = .30$ between studies. This finding is curious as the SST comprehension questions were not edited between studies, and findings from Study 1 did not strongly suggest comprehension questions required the use of ToM abilities. This finding also brings into question the utility of the SST comprehension subscale as a measure of story comprehension. However, given the stark contrast in findings between studies, this observation requires replication.

As previously highlighted, an important characteristic of the SST and SST-B, if they are to hold utility in the measurement of individual differences in neurotypical adults, is their ability to display an absence of ceiling effects and produce varied scores. Aligning with findings from Study 1, both criteria were met in this study. Neither measure displayed ceiling effects, and clear variation was observed across the scores received on the SST and SST-B ToM subscales.

Also of interest was the relationship between performance on the comprehension subscales of the SST and SST-B and their respective ToM subscales. Aligning with Study 1, significant positive correlations emerged for both measures, suggesting that ToM subscale scores were related to comprehension scores. However, consistent with Study 1, findings during Study 2 highlighted limitations in the validity of the comprehension and ToM subscales of the SST and SST-B. Therefore, the ability to make further meaningful interpretations about and inferences from this finding is limited.

Examining the Utility of Simple Self-Report Scales of ToM Ability and Story Comprehension

The purpose of the analyses of single-item measures of ToM and story comprehension was to examine whether simple self-report indices of ToM ability or story comprehension could be substituted for related ability tests due to a growing interest in reducing measure administration times (Yates & Taub, 2003) and to examine the degree to which individuals have metacognitive knowledge about their own ToM skills. However, analyses must be interpreted considering the outlined psychometric properties of the SST and SST-B. Namely, there is limited evidence to support the SST as reflecting ToM within this study. Similarly, the SST-B, while plausibly being incrementally better than the SST in reflecting ToM, also appears to have limited utility for measuring ToM. In taking a cautious approach to interpreting these analyses, findings based on the SST will not be interpreted. The greatest

emphasis will be placed upon the RMET, with some limited consideration given to the SST-B. However, it is acknowledged that there are some concerns regarding the RMET's validity (Higgins et al., 2022), and this study did not thoroughly explore the psychometric properties of the RMET.

Spearman correlations between self-reported ToM ability and performance on the RMET and SST-B were non-significant. While the correlation between the RMET and the self-report scale was above the preregistered criteria of interest ($r_s \geq .10$), this relationship was objectively weak ($r_s = .15$). Given the strength of these correlations, as well as the lack of statistical significance, this finding does not support individuals as having metacognitive knowledge about their own ToM skills and suggests single-item self-report scales are unlikely to be an adequate substitute for the administration of either measure.

This may have been due to the simplicity of the self-report scale, which required participants to rate their perceived ToM ability from one to seven. This scale was likely insufficient to adequately capture the complexity of ToM as a construct. Supporting this, more complex self-report ToM scales have been developed (e.g., The Theory of Mind Inventory-Second Edition, Self-Report; Theory of Mind Inventory: Self Report—Adult) with preliminary psychometric analyses supporting their criterion validity and internal consistency (Crehan et al., 2020), as well as their ability for detecting social cognitive dysfunction (Hutchins et al., 2021). However, these self-report measures are often lengthy (e.g., 48 to 60 items) and use complex rating scales. Thereby, these measures likely hold little utility above commonly used tests of ToM ability in terms of administration time. However, they may still hold utility relative to tests of ToM ability by reducing experimenter burden through their relative ease of administration. Thus, future research may benefit from looking to reduce the number of items on these existing self-report measures to reduce administration time relative to tests of ToM ability.

Regarding story comprehension, analyses suggested that self-reported story understanding was not related to performance on the SST-B comprehension subscale. The absence of a significant relationship may be explainable by findings across Studies 1 and 2, which highlighted that the SST-B comprehension subscale may be an inadequate measure of comprehension or reflecting constructs other than purely story comprehension. Interestingly, the SST comprehension subscale showed a weak but significant positive correlation with self-reported story understanding. This occurred despite the SST comprehension subscale also displaying a significant positive correlation with the RMET in this study. Regardless, this provides some weak support for substituting the SST comprehension subscale with a self-report measurement of story comprehension. However, given the strength of this correlation and other outlined limitations, it is recommended that future researchers do not adopt single-item measures of story comprehension.

These findings provide limited support for substituting ToM measures, or the comprehension subscales of the SST and SST-B, with single-item self-report scales. Likely, this reflects the simplicity of the adopted self-report scales, as more comprehensive self-report measurement tools have shown some utility in reflecting ToM. In light of this, it is recommended that future researchers explore pathways to reduce the administration time of existing self-report measures (e.g., through identification and omission of items that are poorly indexing ToM) as opposed to adopting the self-report scales used in this study.

Examining whether Familiarity with Literary Fiction is Associated with Greater ToM Abilities

When drawing conclusions regarding ToM, the greatest emphasis will be placed on analyses conducted utilising the RMET to remain consistent with earlier identified limitations in the psychometric properties of the measures employed. Some cautious interpretation of findings utilising the SST-B will also be undertaken.

In line with anticipated findings, the RMET and SST-B displayed a significant positive correlation with the Author Recognition Test (ART), a measure of familiarity with literary fiction. This finding aligns with wider literature and replicates the finding identified in Mumper and Gerrig's (2017) meta-analysis. However, the effect sizes observed in the present study (r_s = .32 - .41) are notably larger than those identified by Mumper and Gerrig for correlational studies (r = .168). These differences are unlikely to be related to the ToM measure adopted, as four of the five studies in Mumper and Gerrig's meta-analysis utilised the RMET. However, all five studies used an alternative version of the ART, the ART-Revised (ART-R; Mar et al., 2006). The ART-R was developed to examine whether familiarity with authors of non-fiction versus fiction was associated with differences in cognitive abilities. The measure included a fiction and non-fiction subscale, each with 50 authors and 40 distractors. Of note, while the ART-R shares some overlap in items with the ART, the 50 retained fiction authors and 40 distractors does notably differ from the 65 authors and 65 distractors used in the present study. As such, differences in the measures of literary fiction familiarity used across studies may partially account for the notably larger observed effect sizes.

Cultural or sample characteristics may also be implicated in the observed differences between the present study and the wider literature. Namely, most of the studies analysed in Mumper and Gerrig's (2017) meta-analysis used samples of young (i.e., with the exception of one study, the mean ages ranged from 18.9 to 22.3) female participants from the United States or Canada. Alternatively, the present study was relatively older (i.e., the mean age in the present study was 33.3), primarily male, and used participants from New Zealand, Australia, Canada, the United States, and the United Kingdom. Given that culture (Oi et al., 2013), age (Henry et al., 2013), and sex (Baron-Cohen et al., 2022), all influence ToM

abilities, these variables may also be partially contributing to the observed differences between studies.

The SST also showed a significant positive correlation with the ART, despite findings suggesting that this measure is not sufficiently indexing ToM. Plausibly, individuals who score high on the ART may have greater proficiency in some other ability (e.g., language) which these measures may all indirectly be indexing. Supporting this assertion, it is widely acknowledged that supposed ToM measures commonly require understanding complex language (San José Cáceres et al., 2014). There is emerging evidence to suggest that language attenuates the relationship between autism disorder symptomology and performance on ToM measures (Gernsbacher & Yergeau, 2019). Thus, the plausibility that the relationships between the ART and the SST, SST-B, and RMET are being attenuated by language ability is not unfounded. However, with the available data, further exploration regarding a third variable explanation is not possible and should be the focus of future research.

Summary and Future Directions

The primary findings are summarised below in reference to the overarching aims for Study 2 outlined in Chapter 4 and Chapter 8:

1. *To determine whether the Short Story Task, or the alternate form piloted during Study 1, is a better measure of ToM with neurotypical adults.*
    - *1.1. To determine whether self-report scales of ToM ability and short story comprehension are adequate substitutes for tests of ToM ability and short story comprehension.*

Evidence suggested that when delivered in their originally intended forms, the SST-B may be an incrementally more valid measure of ToM relative to the SST, aligning with findings from Study 1. However, both the SST and SST-B require refinement and further

examination of their psychometric properties before they are adopted in future research as measures of ToM. Further, adjustments are also needed to the marking rubrics of the SST and SST-B to improve consistency in score application. Regarding Aim 1.1, the self-report scales adopted in the present study were inadequate substitutes for performance on purported measures of ToM and story comprehension.

2. *Replicate the finding that familiarity with literary fiction is associated with greater ToM abilities (Mumper & Gerrig, 2017).*

Findings indicated evidence of a positive relationship between performance on the RMET/SST-B and a measure of literary fiction familiarity, replicating Mumper and Gerrig's (2017) findings. The observed relationship may have been stronger than Mumper and Gerrig's findings due to demographic differences between studies and the use of the ART instead of the ART-R.

Several avenues for further research have been identified and articulated throughout this chapter. Undoubtedly, the primary avenue for future research is to refine ToM measures such as the SST and SST-B to improve their validity. Refinement may plausibly involve editing the items in the scales, changes in the delivery of the measure, and, in line with recommendations outlined by Beaudoin et al. (2020), clearer delineation regarding the specific subprocesses the measures are designed to index (e.g., emotional attribution, intention of action/motivation). Following refinement, establishment, and ongoing replication of the psychometric properties of these scales, is also essential. These ideas will be explored in greater depth within the General Discussion.

CHAPTER 12:

GENERAL DISCUSSION

Summary of Findings

This thesis initially aimed to explore three areas: the relationship between video game play and Theory of Mind (ToM); how adequate existing media psychology theories were for investigating changes in ToM; and whether a newly constructed alternate form of the Short Story Task (SST) had adequate psychometric properties. In order to achieve these aims, Study 1 sought to validate an alternative form of the SST. Initial psychometric analyses of the SST in Study 1 brought into question its validity as a measure of ToM. Alternatively, the newly constructed version of the SST, the Short Story Task-B (SST-B), showed some evidence of being an incremental improvement over the SST in its ability to reflect ToM through a stronger positive relationship with the Reading the Mind in the Eyes Test – Revised Edition (RMET). As such, exploratory analyses examining relationships between video game play and ToM were conducted using the RMET and the SST-B.

Exploratory analyses indicated the presence of a negative relationship between engagement with fighting and action video games and performance on the RMET. When controlling for time, a similar negative relationship emerged between the SST-B and fighting game engagement. This finding provided some support for the General Aggression Model (GAM) but not the General Learning Model (GLM). However, negative correlations did not emerge between the RMET/SST-B and other game genres, which likely included violence, suggesting either the observed findings may have occurred by chance, or the GAM has limitations in its utility within this area. Analyses examining the relationship between the social context of video game play and ToM did not align with predictions made by the GLM, further highlighting gaps in the utility of existing media psychology theory. Overall, findings from Study 1 indicated a) limited evidence of a relationship between video game genre

engagement and purported measures of various ToM abilities and b) existing video game theories were generally inadequate for making predictions about ToM abilities.

While Study 1 highlighted that the SST and SST-B were poor alternate forms, Study 2 primarily investigated whether the SST-B was an incremental improvement to the measurement of ToM relative to the SST. While findings aligned with Study 1 and suggested that the SST-B may be an improvement on the SST, these improvements were modest, and the findings also highlighted that both measures, when delivered in their originally intended forms, appeared to be generally unfit for purpose. Thus, this chapter will focus on the wider implications of this finding, why these measures might have questionable validity, and thereby, the nature of the constructs they may be reflecting, how threats to the validity of these measures may plausibly be addressed, and future directions for research.

Implications for ToM Measurement

The findings of Studies 1 and 2 have implications for the ongoing use of the SST and SST-B, as well as ToM measurement generally. Regarding the SST and SST-B, these findings highlight that these measures should be refined before they are utilized in further research. Research adopting these measures in their current forms may produce invalid or contradictory findings and are likely to contribute further to the conceptual ambiguity which is prevalent within the ToM literature (François & Rossetti, 2020). Relatedly, these findings bring into question conclusions drawn based upon scores derived from the SST in previous research (e.g., findings by Trott and Bergen (2020) regarding the relationship between mentalization, as measured by the SST, and pragmatic inferences). Thus, replication of previous findings drawn based upon these measures is recommended following measure refinement.

While this thesis primarily focused on the SST and SST-B, these findings plausibly have implications for many ToM measures. Specifically, psychometric evaluations of other ToM measures often produce contradictory findings similar to those observed in this thesis (Gernsbacher & Yergeau, 2019). Highlighting this, the RMET has been subject to several confirmatory factor evaluations examining single and three-factor models (Benau et al., 2020; Black, 2019; Higgins et al., 2022; Olderbak et al., 2015; Sherman et al., 2015). These findings have drawn contrasting conclusions, stating that a single-factor model is plausible (Benau et al., 2020; Black, 2019; Sherman et al., 2015), while others have found limited support for single and three-factor models (Higgins et al., 2022; Olderbak et al., 2015). These contrasting findings have, in part, contributed to ongoing speculation regarding the validity of this measure. Thereby, although the present chapter focuses on the SST and SST-B, the conclusions and recommendations throughout this chapter may plausibly have implications for other measures of ToM that remain contested.

While it has been established that the SST and SST-B, like many other ToM measures, have questionable validity, this conclusion begs the question: If these measures are inadequately indexing ToM, what, precisely, are they measuring? Within this study, there is insufficient information to draw definitive conclusions in this area. However, existing literature may offer some insight. Based on these speculations, future hypotheses can be generated and investigated to improve accuracy in the measurement of ToM.

What Might the SST/SST-B and Other ToM Measures be Reflecting, and how can this be Rectified?

Questions about measure validity are an area of longstanding contention in the ToM literature. Within this space, several bodies of literature warrant consideration when speculating regarding what the SST, SST-B, and other purported ToM measures, may be indexing; the need for definitional clarity in ToM measurement; measurement of lower-level

cognitive processes; use of compensatory strategies; and co-activation of higher-order cognitive processes. Thankfully, as detailed later in this chapter, there are amendments that can be made to the SST and SST-B to plausibly address the majority of these shortcomings and improve their validity. Future research aiming to amend these measures and re-examine their psychometric properties would benefit from addressing these threats to their validity in an iterative process. This process would allow for clear identification of which changes improve their validity and should be retained in future administrations. While this discussion focuses on the SST and SST-B, other researchers looking to amend other ToM measures may benefit from adopting some of the outlined recommendations.

Definitional Ambiguity. The questionable validity of many ToM measures, including the SST and SST-B, may partly stem from how ToM is currently defined. It has long been recognized that ToM is a complex ability that a single measure is unlikely to capture, leading to the investigation and articulation of widely supported subordinate processes that may underpin ToM (e.g., affective and cognitive ToM systems; Shamay-Tsoory et al., 2007). Thus, the use of multiple ToM measures is often recommended when one wants to more comprehensively investigate an individual's ToM abilities (Smogorzewska et al., 2018).

More recently, it has been suggested that these dual-process theories may still be too simplistic to reflect the complexity of ToM accurately. This has led some researchers to recommend that theories of ToM subprocesses require further divisions and ToM measures require more specific definitions (Warnell & Redcay, 2019). For example, supposed affective ToM measures may require a person to understand someone's emotional state following a specific situation, that different people can have different emotional reactions to the same situation, and that some people may hide their emotions. Alternatively, cognitive ToM measures may require a person to understand that different people can achieve the same result with different motivations, other people can hold different beliefs, or that people can hold a

range of different and often conflicting desires (Beaudoin et al., 2020). This list is far from exhaustive but highlights how measures that are supposedly indexing the same construct or subprocesses may be examining radically different and only distally related sub-abilities of ToM.

Unsurprisingly, when examining the relationship between two ToM measures which should theoretically be related, a relationship may not emerge (Ahmed & Miller, 2011; Warnell & Redcay, 2019). This may lead to questions regarding the validity of one or both measures used, despite them still plausibly being valid measures of ToM sub-abilities. By clearly defining what sub-abilities existing ToM measures are supposedly reflecting, researchers can draw more informed theory regarding which measures are likely to be related or unrelated to one another. This would allow for more thorough and informed investigations of the validity of ToM measures. Thus, current limitations in how ToM is defined and operationalized may contribute to the questionable validity of the SST and SST-B despite the possibility these measures may still plausibly reflect some sub-ability/abilities of ToM.

To address this limitation, articulating the supposed sub-abilities the SST and SST-B are indexing is a necessary initial step. This task is complicated, as the SST and SST-B were designed to examine a range of different mental states, including beliefs, feelings, and intentions (Dodell-Feder et al., 2013). This is not uncommon, with Beaudoin et al. (2020) identifying that at least 11.4% of ToM measures designed for children reflected multiple sub-abilities of ToM. However, based on ToM sub-ability categories articulated by Beaudoin et al., it is likely that the SST and SST-B should primarily reflect intention explanations (explaining individuals' intentions/motivations for actions within a social situation) and how desires influence emotions and actions. While these two sub-abilities may be the primary abilities the SST and SST-B are reflecting; they are also likely tapping a range of other sub-abilities to a lesser degree. For example, question five on the SST requires the identification

of sarcasm and would reflect the sub-ability of sarcasm/irony. Thus, for measures designed to reflect more than one sub-ability, there may be utility in delineating between primary and secondary sub-abilities that are theoretically indexed.

Unfortunately, in isolation, identification of the sub-abilities that the SST and SST-B may index is insufficient to overcome the threat of definitional ambiguity to their validity. This limitation may potentially be overcome through the hybridization of theoretical models/frameworks outlined by Beaudoin et al. (2020) and Schaafsma et al. (2015), whereby all ToM measures are broken down into their theoretical sub-abilities, and then theoretical relationships between these sub-abilities are articulated. This sort of theory/model building is an extensive undertaking that is beyond the scope of the present thesis. However, creating such a theoretical framework would provide future researchers with a strong theoretical foundation to draw subsequent hypotheses and test the psychometric properties of ToM measures.

Reflection of Lower-Level Processes. While definitional opacity may be contributing to the apparent questionable validity of ToM measures, other authors argue that many ToM measures fail to meet the basic criteria for what constitutes ToM measurement. François and Rosetti (2020) assert that irrespective of what sub-abilities ToM may rely upon, the minimum requirement for a ToM measure to be valid is: 1) the necessitation of mental state representation, and 2) differentiation between self and others' mental states. Measures that fail to meet these criteria are contended to instead quantify lower-level cognitive processes given the centrality of mental state representation to the ToM construct. For example, François and Rosetti claim that visual emotion ascription measures (e.g., the RMET) fail to meet criteria 1) as no mental state representation is required for task success. As such, they allege the RMET is a measure of facial expression discrimination, not ToM.

Regarding the SST and SST-B, both measures include questions that meet criteria 1) and 2) as they explicitly ask individuals to determine the mental states that are motivating characters to act in particular ways (e.g., Why is Nick afraid to look at Marjorie?; Why does Pinin look at the floor when the major asks him "And you really don't want-" and "That your great desire isn't really-"?). However, both measures also include questions which, arguably, do not require mental state attribution (e.g., Who is Bill and what does he reveal when he asks Nick, "Did she go alright? … Have a scene?"?**;** Who is James and what does he reveal when he asks, "So the major propositioned you too?"). In this example, determining who Bill or James is requires the participant to use contextual story clues to determine the relationship between the protagonist and these individuals, a process that does not require attributing a mental state to any character. These questions may instead reflect abilities other than ToM (e.g., comprehension), suggesting that both the SST and SST-B may partially index lower-level cognitive processes than ToM.

This limitation highlights a clear pathway to plausibly improving the validity of the SST, SST-B, and other measures. Across the SST and SST-B, all items which fail to meet these two criteria could be edited or potentially omitted. This process may also involve omitting items identified through factor analyses as poorly indexing the latent construct. Additional items which meet criteria 1 and 2 outlined above may need to be generated to ensure ceiling effects do not occur across the measures. Subsequently, psychometric analyses may be reconducted to determine whether the deletion of these items and the potential generation of new items has improved the measures' validities.

Compensation. The use of compensatory strategies, an area of increasing interest within the autism spectrum disorder and ToM literature (Livingston et al., 2019), may provide further insight regarding what the SST and SST-B may be reflecting. Generally, compensation allows individuals with ToM deficits to mask associated behaviours using

other behavioural or cognitive strategies/abilities (e.g., rote learning and enacting social rules; using memory, executive functions, or intelligence abilities to mitigate ToM deficits). These compensation strategies may theoretically be used to bolster performance on measures of ToM (Livingston & Happé, 2017). However, further research is required in this area, given that there has been limited investigation into how cognitive compensation may improve performance on ToM tasks (Livingston et al., 2019).

Despite this being a new and emerging area of research, compensation is a clear theoretical threat to the validity of many ToM measures, with authors speculating that compensation may play an explanatory role in the contrasting findings across validity analyses of many ToM measures (Higgins et al., 2022). Importantly in the context of this study, there is emerging evidence to suggest that compensatory strategies are also utilised by neurotypical adults, albeit to a lesser extent than those with autism spectrum disorder (Livingston et al., 2020). Thus, while the mechanisms underpinning cognitive compensation in the context of ToM assessment remain unclear, the SST and SST-B may inadvertently be indexing other cognitive abilities that individuals employ to improve performance.

Given that compensation in ToM is a relatively new area of investigation, few researchers have explored avenues to mitigate individuals utilising compensatory strategies in ToM assessment. However, there is evidence to suggest that compensatory strategies are likely time-intensive (Miu et al., 2012). By imposing deadlines on the SST and SST-B (e.g., only providing 15 minutes to complete the measure or only providing 30 seconds per question), participants may be unable to effectively utilise most compensatory strategies, thereby improving the validity of these measures. Plausibly, the inclusion of a deadline may also improve the ecological validity of ToM measures, as real-life dynamic social situations necessitate ToM systems to work rapidly and under pressure.

Co-Activation of Higher-Order Abilities. Relatedly, tests of ToM often strongly correlate with a range of higher-order cognitive abilities. For example, ToM tests have been found to correlate with tests of memory (Crane et al., 2013), executive function (Ahmed & Miller, 2011; Lecce et al., 2017), and language (Im-Bolter et al., 2016). Troublingly, while evidence is mixed, there is some suggestion that individual differences in these other cognitive abilities influences performance on ToM tasks (Laillier et al., 2019).

Exemplifying the possible confounding impact of other cognitive abilities on ToM performance is emerging evidence challenging the ToM deficit hypothesis in autism spectrum disorder. Specifically, some researchers suggest that communication deficits that characterize autism spectrum disorder contribute to impaired performance on linguistically complex ToM tasks, as opposed to true deficits in ToM ability (Gernsbacher & Yergeau, 2019). While further research is needed to support this claim, it aligns with the lived experience of individuals with autism spectrum disorder (Holt et al., 2022) and research highlighting that ToM and language abilities are often difficult to differentiate (Balaban et al., 2016). Such findings have wide-reaching implications for the measurement of ToM. Notably, they cast further doubt on the validity of existing ToM measures and suggest that using measures of autism spectrum disorder symptomology in ToM validity assessments may be inappropriate.

While the research described above emphasizes the confounding impact of other cognitive abilities in the context of autism spectrum disorder, it is likely that other cognitive abilities also influence ToM performance in neurotypical adults (Laillier et al., 2019). This suggests that performance on the SST and SST-B may also be inadvertently influenced by an individual's other cognitive abilities and not necessarily ToM. Minimizing or mitigating the impact that other cognitive abilities may have on the validity of the SST, SST-B, or other ToM measures is difficult, given that ToM is intrinsically intertwined with other cognitive

systems (Bivona et al., 2018). However, future psychometric evaluations of these measures would benefit from adopting a battery of cognitive and neuropsychological tests to determine what abilities the SST and SST-B are related to and whether performance on these other tests influences performance on the SST and SST-B. While this does not directly improve the validity of these measures, it may offer some insight into other abilities they may be indexing. This would thereby allow future researchers to measure and control for the influence of these other abilities or to refine the SST and SST-B further, so item content is less reflective of other cognitive systems.

Future Directions

As a priority, future studies should look to 1) reformulate existing ToM frameworks to improve specificity in the constructs that ToM measures may be indexing, 2) refine and synthesize existing ToM measures to improve their likelihood of reflecting ToM, as opposed to lower-level cognitive processes, 3) examine the utility of deadlines in mitigating the impact of compensation on ToM measure performance, and 4) determine the impact of higher-order cognitive abilities in ToM measure performance, and either control for these abilities or further refine ToM measures. Given that the accurate measurement of ToM is a prerequisite to any research looking to explore ToM, research investigating these four areas outlined above is a logical first step.

Should the validity of the SST, SST-B, and other ToM measures be improved through the recommendations mentioned above, the future of ToM research is vast. In the context of this thesis, the findings from Study 1's video game analyses would benefit from replication, given that drawn conclusions are limited by the ToM measures used. These findings could be expanded upon through subsequent longitudinal or experimental methodologies, as this thesis initially intended to do. Similarly, findings examining relationships between ToM abilities and literary fiction familiarity found within Study 2 and by Mumper and Gerrig (2017),

would also benefit from replication, given that these findings were also limited by the ToM measures employed. Subsequent research may then look to replicate many findings across the ToM literature, which have also potentially been constrained by limitations in measure validity that this thesis has highlighted.

Limitations

The present work is not without limitations. For instance, a limitation of Study 2 was the sample size employed. The required sample size was calculated utilizing power analysis. While this provided a sample size that was appropriate for the majority of statistical procedures employed, this process did not adequately consider the required sample size needed for conducting the confirmatory factor analyses. This limited the degree to which resultant findings could be meaningfully interpreted. It is recommended that future researchers conducting factor analyses consider a range of variables in calculating appropriate sample sizes, such as the method of estimation and the number of statistical estimates within the hypothesized model.

Similarly, another limitation across studies was departures from preregistered protocol. While these departures were necessary to conduct meaningful analyses and interpret resultant findings, it is acknowledged that these deviations weaken the strength of drawn conclusions. Regarding Study 2, departures were notable regarding confirmatory factor analyses. In addition to adopting widely accepted cut-off values, a wide array of fit indices were reported to mitigate concerns regarding Type I error (Hussey & Hughes, 2020). Further, findings were interpreted with caution. It is hoped that highlighting this limitation will serve as a reminder to future researchers regarding the importance of thoroughly considering data analysis procedures during the preregistration phase of research.

It also requires acknowledgement that the employed method, and resultant validity analyses, meant that the validity of the SST and SST-B were not comprehensively assessed (e.g., convergent and discriminant validity of these measures were not examined throughout either study). While these facets of validity have been partially examined in previous research for the SST (Dodell-Feder et al., 2013; Giordano et al., 2019), the failure to observe sufficiently meaningful correlations between the RMET and SST across Studies 1 and 2 necessitates further replication of these findings. Within this thesis, this limitation primarily arose due to concerns regarding survey length, as the median completion time for Study 1 was 49 minutes, and for Study 2 was 57 minutes and 36 seconds. This was of concern, as increasing survey length can lead to participant fatigue, thereby increasing measurement error (Egleston et al., 2011). Should future researchers look to explore the psychometric properties of the SST or SST-B, it is recommended that they reduce survey length through an independent examination of these measures. Concurrently, it is recommended that these researchers implement other measures (e.g., The Interpersonal Reactivity Index; Davis, 1983) to further examine the validity of the SST or SST-B.

Findings from this thesis partially rested on the presumption that the RMET was likely reflecting some component of ToM given that 1) limited other measures were available to potentially reflect the ToM abilities of neurotypical adults, 2) the psychometric properties of the RMET were not explored in this thesis, 3) there is existing literature that supports the psychometric properties of the RMET (F. J. Ferguson & Austin, 2010; Vellante et al., 2013), and 4) the RMET is ongoingly used as a measure of ToM in current literature (e.g., Lee et al., (2021)). However, our findings align with wider literature casting doubt on the validity of existing ToM measures (Bosco et al., 2016), from which the RMET is likely not exempt. While these limitations were acknowledged throughout this thesis, reflection is required on

how findings throughout this thesis may be impacted or limited should the RMET not be adequately indexing ToM.

Researchers who assert that the RMET may not be indexing ToM almost unanimously agree that the RMET is likely instead reflecting some form of emotion recognition (François & Rossetti, 2020; Higgins et al., 2022; Olderbak et al., 2015). While a distinct construct, emotion recognition and ToM are posited to be closely related (Mier et al., 2010). A positive relationship would still be expected between purported measures of ToM (e.g., the SST and SST-B) and the RMET. Thereby, the observation of a positive relationship between the RMET and SST-B, but not the SST, would still provide some support for the validity of the SST-B. However, the strength of this validity evidence would be slightly tempered given the relationship would be between two theoretically related but distinct constructs.

Regarding the use of the RMET in the video game analyses conducted during Study 1, the majority of observed findings would still stand. Specifically, existing theories still appear inadequate to account for observed findings. However, should the RMET be found to be a measure of emotion recognition, it would be inappropriate to draw conclusions between video game play and ToM. However, the comparisons made to the wider literature are still warranted, given that the two identified studies that examined relationships between video games and ToM adopted the RMET (Bormann & Greitemeyer, 2015; Kühn et al., 2019).

Concluding Remarks

This thesis initially examined whether video game play and social context were related to ToM and whether the GAM or GLM better accounted for observed findings. In addition, the SST-B was created and piloted to address the absence of alternate forms of existing ToM measures that could be used with neurotypical adults. Findings from Study 1 highlighted limitations in the validity of both the SST and SST-B but suggested that the SST-

B may be an incrementally better measure of ToM. Future research was compelled to further investigate the psychometric properties of the SST and SST-B.

As such, a second study further explored the psychometric properties of the SST and SST-B. Findings from Study 1 were largely replicated, but confirmatory factor analyses highlighted that the SST and SST-B were generally unfit for purpose when delivered in their current forms. This conclusion fits with wider literature that is beginning to question how ToM is defined (François & Rossetti, 2020) and how valid current tools are to measure ToM (Higgins et al., 2022). Plausible threats to the validity of the SST and SST-B were thereby considered, such as the need for definitional clarity in ToM measurement; measurement of lower-level cognitive processes; use of compensatory strategies; and co-activation of higher-order cognitive processes. Plausible ways to rectify limitations in the validity of the SST and SST-B were also outlined, including a reformulation of existing ToM frameworks; editing or deletion of items that may quantify lower-level processes; using deadlines to mitigate compensation; and further refinement of ToM measures to reduce their reliance on other higher-order cognitive processes. Future research is compelled to address these limitations, given their wide-reaching implications for the ToM literature.

REFERENCES

Acheson, D. J., Wells, J. B., & MacDonald, M. C. (2008). New and updated tests of print exposure and reading abilities in college students. *Behavior Research Methods, 40*(1), 278–289. https://doi.org/10.3758/BRM.40.1.278

Adams, E. (2014). *Fundamentals of game design*. Pearson Education.

Adelantado-Renau, M., Moliner-Urdiales, D., Cavero-Redondo, I., Beltran-Valls, M. R., Martínez-Vizcaíno, V., & Álvarez-Bueno, C. (2019). Association between screen media use and academic performance among children and adolescents: A systematic review and meta-analysis. *JAMA Pediatrics, 173*(11), 1058–1067. https://doi.org/10.1001/jamapediatrics.2019.3176

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

Allen, J. J., Anderson, C. A., & Bushman, B. J. (2018). The general aggression model. *Current Opinion in Psychology, 19*, 75–80. https://doi.org/10.1016/j.copsyc.2017.03.034

American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders* (5th ed.). American Psychiatric Publishing.

Anderson, C. A., & Bushman, B. J. (2002). Human aggression. *Annual Review of Psychology, 53*, 27–51. https://doi.org/10.1146/annurev.psych.53.100901.135231

Anderson, C. A., & Bushman, B. J. (2018). Media violence and the general aggression model. *Journal of Social Issues, 74*(2), 386–413. https://doi.org/10.1111/josi.12275

Anderson, C. A., & Carnagey, N. L. (2004). Violent evil and the general aggression model. In A. G. Miller (Ed.), *The social psychology of good and evil* (pp. 168–192). The Guilford Press.

Anderson, C. A., Shibuya, A., Ihori, N., Swing, E. L., Bushman, B. J., Sakamoto, A., Rothstein, H. R., & Saleem, M. (2010). Violent video game effects on aggression, empathy, and prosocial behavior in Eastern and Western countries: A meta-analytic review. *Psychological Bulletin, 136*(2), 151–173. https://doi.org/10.1037/a0018251

Apperly, I. A. (2008). Beyond simulation–theory and theory–theory: Why social cognitive neuroscience should use its own concepts to study "Theory of Mind." *Cognition, 107*(1), 266–283. https://doi.org/10.1016/j.cognition.2007.07.019

Atkinson, D., & Murray, M. (1987). *Improving Interrater Reliability.* College Composition and Communication, Atlanta, Georgia.

Bailey, K., & West, R. (2013). The effects of an action video game on visual and affective information processing. *Brain Research, 1504*, 35–46. http://dx.doi.org/10.1016/j.brainres.2013.02.019

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, 44*, 78–92. https://doi.org/10.1016/j.intell.2014.03.001

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

Barash, D. P., & Lipton, J. E. (2011). *Payback: Why we retaliate, redirect aggression, and take revenge*. Oxford University Press.

Barlett, C. P., Anderson, C. A., & Swing, E. L. (2009). Video game effects—Confirmed, suspected, and speculative: A review of the evidence. *Simulation & Gaming*, *40*(3), 377–403. http://dx.doi.org/10.1177/1046878108327539

Barlett, C. P., Harris, R. J., & Bruey, C. (2008). The effect of the amount of blood in a violent video game on aggression, hostility, and arousal. *Journal of Experimental Social Psychology*, *44*(3), 539–546. http://dx.doi.org/10.1016/j.jesp.2007.10.003

Baron-Cohen, S., O'riordan, M., Stone, V., Jones, R., & Plaisted, K. (1999). Recognition of faux pas by normally developing children and children with Asperger syndrome or high-functioning autism. *Journal of Autism and Developmental Disorders*, *29*(5), 407–418. https://doi.org/10.1023/a:1023035012436

Baron-Cohen, S., Radecki, M. A., Greenberg, D. M., Warrier, V., Holt, R. J., & Allison, C. (2022). Sex differences in theory of mind: The on-average female advantage on the Reading the Mind in the Eyes Test. *Developmental Medicine and Child Neurology*, *64*(12), 1440–1441. https://doi.org/10.1111/dmcn.15364

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

Barone, P., Corradi, G., & Gomila, A. (2019). Infants' performance in spontaneous-response false belief tasks: A review and meta-analysis. *Infant Behavior and Development*, *57*, Article 101350. https://doi.org/10.1016/j.infbeh.2019.101350

Basak, C., Boot, W. R., Voss, M. W., & Kramer, A. F. (2008). Can training in a real-time strategy video game attenuate cognitive decline in older adults? *Psychology and Aging*, *23*(4), 765–777. https://doi.org/10.1037%2Fa0013494

Beaudoin, C., Leblanc, É., Gagner, C., & Beauchamp, M. H. (2020). Systematic review and inventory of theory of mind measures for young children. *Frontiers in Psychology*, *10*, Article 2905. https://doi.org/10.3389/fpsyg.2019.02905

Bediou, B., Adams, D. M., Mayer, R. E., Tipton, E., Green, C. S., & Bavelier, D. (2018). Meta-analysis of action video game impact on perceptual, attentional, and cognitive skills. *Psychological Bulletin*, *144*(1), 77–110. https://doi.org/10.1037/bul0000130

Bejanin, A., Chételat, G., Laisney, M., Pélerin, A., Landeau, B., Merck, C., Belliard, S., de La Sayette, V., Eustache, F., & Desgranges, B. (2017). Distinct neural substrates of affective and cognitive theory of mind impairment in semantic dementia. *Social Neuroscience*, *12*(3), 287–302. https://doi.org/10.1080/17470919.2016.1168314

Belchior, P., Marsiske, M., Sisco, S. M., Yam, A., Bavelier, D., Ball, K., & Mann, W. C. (2013). Video game training to improve selective visual attention in older adults. *Computers in Human Behavior*, *29*(4), 1318–1324. https://doi.org/10.1016/j.chb.2013.01.034

Benau, E. M., Wiatrowski, R., & Timko, C. A. (2020). Difficulties in emotion regulation, alexithymia, and social phobia are associated with disordered eating in male and female undergraduate athletes. *Frontiers in Psychology*, *11*, Article 1646. https://doi.org/10.3389/fpsyg.2020.01646

Bivona, U., Formisano, R., Mastrilli, L., Zabberoni, S., Caltagirone, C., & Costa, A. (2018). Theory of mind after severe acquired brain injury: Clues for interpretation. *BioMed Research International*, *2018*, 1–12. https://doi.org/10.1155/2018/5205642

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

Blocker, K. A., Wright, T. J., & Boot, W. R. (2014). Gaming preferences of aging generations. *Gerontechnology: International Journal on the Fundamental Aspects of Technology to Serve the Ageing Society*, *12*(3), 174–184. https://doi.org/10.4017%2Fgt.2014.12.3.008.00

Boot, W. R., Kramer, A. F., Simons, D. J., Fabiani, M., & Gratton, G. (2008). The effects of video game playing on attention, memory, and executive control. *Acta Psychologica*, *129*(3), 387–398. https://doi.org/10.1016/j.actpsy.2008.09.005

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

Borgonovi, F. (2016). Video gaming and gender differences in digital and printed reading performance among 15-year-olds students in 26 countries. *Journal of Adolescence*, *48*, 45–61. https://doi.org/10.1016/j.adolescence.2016.01.004

Bormann, D., & Greitemeyer, T. (2015). Immersed in virtual worlds and minds: Effects of in-game storytelling on immersion, need satisfaction, and affective theory of mind. *Social Psychological and Personality Science*, *6*(6), 646–652. https://doi.org/10.1177/1948550615578177

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

Bottiroli, S., Cavallini, E., Ceccato, I., Vecchi, T., & Lecce, S. (2016). Theory of Mind in aging: Comparing cognitive and affective components in the faux pas test. *Archives of Gerontology and Geriatrics*, *62*, 152–162. https://doi.org/10.1016/j.archger.2015.09.009

Bowen, N. K., & Guo, S. (2011). *Structural equation modeling*. Oxford University Press.

Brand, J. E., Todhunter, S., & Jervis, J. (2017). *Digital New Zealand 2018*. IGEA.

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

Brunet, E., Sarfati, Y., Hardy-Baylé, M.-C., & Decety, J. (2000). A PET investigation of the attribution of intentions with a nonverbal task. *Neuroimage*, *11*(2), 157–166. https://doi.org/10.1006/nimg.1999.0525

Buckley, D., Codina, C., Bhardwaj, P., & Pascalis, O. (2010). Action video game players and deaf observers have larger Goldmann visual fields. *Vision Research*, *50*(5), 548–556. https://doi.org/10.1016/j.visres.2009.11.018

Buckley, K. E., & Anderson, C. A. (2006). A theoretical model of the effects and consequences of playing video games. In P. Vorderer & J. Bryant (Eds.), *Playing video games: Motives, responses, and consequences* (pp. 363–378). Lawrence Erlbaum Associates Publishers.

Buelow, M. T., Okdie, B. M., & Cooper, A. B. (2015). The influence of video games on executive functions in college students. *Computers in Human Behavior*, *45*, 228–234. https://doi.org/10.1016/j.chb.2014.12.029

Burnay, J., Kepes, S., & Bushman, B. J. (2022). Effects of violent and nonviolent sexualized media on aggression-related thoughts, feelings, attitudes, and behaviors: A meta-analytic review. *Aggressive Behavior*, *48*(1), 111–136. https://doi.org/10.1002/ab.21998

Cassels, T. G., & Birch, S. A. (2014). Comparisons of an open-ended vs. Forced-choice 'mind reading'task: Implications for measuring perspective-taking and emotion recognition. *PLoS One*, *9*(12), Article e93653. https://doi.org/10.1371/journal.pone.0093653

Castel, A. D., Pratt, J., & Drummond, E. (2005). The effects of action video game experience on the time course of inhibition of return and the efficiency of visual search. *Acta Psychologica*, *119*(2), 217–230. https://doi.org/10.1016/j.actpsy.2005.02.004

Cerniglia, L., Bartolomeo, L., Capobianco, M., Lo Russo, S. L. M., Festucci, F., Tambelli, R., Adriani, W., & Cimino, S. (2019). Intersections and divergences between empathizing and mentalizing: Development, recent advancements by neuroimaging and the future of animal modeling. *Frontiers in Behavioral Neuroscience*, *13*, 1–17. https://doi.org/10.3389/fnbeh.2019.00212

Charles Jr, J. (1995). Hemingway's complicated" Enquiry" in'Men Without Women.'. *Studies in Short Fiction*, *32*(2), 217–223.

Chen, Y.-Q., & Hsieh, S. (2018). The relationship between internet-gaming experience and executive functions measured by virtual environment compared with conventional laboratory multitasks. *PloS One*, *13*(6), Article e0198339. https://doi.org/10.1371/journal.pone.0198339

Clark, L. A., Cuthbert, B., Lewis-Fernández, R., Narrow, W. E., & Reed, G. M. (2017). Three approaches to understanding and classifying mental disorder: ICD-11, DSM-5, and the National Institute of Mental Health's Research Domain Criteria (RDoC). *Psychological Science in the Public Interest*, *18*(2), 72–145. https://doi.org/10.1177/1529100617727266

Clause, C. S., Mullins, M. E., Nee, M. T., Pulakos, E., & Schmitt, N. (1998). Parallel test form development: A procedure for alternate predictors and an example. *Personnel Psychology*, *51*(1), 193–208.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Lawrence Erlbaum Associates.

Cotter, J., Granger, K., Backx, R., Hobbs, M., Looi, C. Y., & Barnett, J. H. (2018). Social cognitive dysfunction as a clinical marker: A systematic review of meta-analyses across 30 clinical conditions. *Neuroscience & Biobehavioral Reviews*, *84*, 92–99. https://doi.org/10.1016/j.neubiorev.2017.11.014

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

Crehan, E. T., Althoff, R. R., Riehl, H., Prelock, P. A., & Hutchins, T. (2020). Brief report: Me, reporting on myself: Preliminary evaluation of the criterion-related validity of the Theory of Mind Inventory-2 when completed by autistic young adults. *Journal of Autism and Developmental Disorders*, *50*(2), 659–664. https://doi.org/10.1007/s10803-019-04278-5

David, L. T. (2012). Training of spatial abilities through computer games–results on the relation between game's task and psychological measures that are used. *Procedia-Social and Behavioral Sciences*, *33*, 323–327. http://dx.doi.org/10.1016/j.sbspro.2012.01.136

Davis, M. H. (1983). Measuring individual differences in empathy: Evidence for a multidimensional approach. *Journal of Personality and Social Psychology*, *44*(1), 113–126. https://doi.org/10.1037/0022-3514.44.1.113

Devilly, G. J., Brown, K., Pickert, I., & O'Donohue, R. (2017). An evolutionary perspective on cooperative behavior in gamers. *Psychology of Popular Media Culture*, *6*(3), 208–221. https://doi.org/10.1037/ppm0000097

Devilly, G. J., O'Donohue, R. P., & Brown, K. (2021). Personality and frustration predict aggression and anger following violent media. *Psychology, Crime & Law*, 1–37. https://doi.org/10.1080/1068316X.2021.1999949

Diaz, R. L., Wong, U., Hodgins, D. C., Chiu, C. G., & Goghari, V. M. (2016). Violent video game players and non-players differ on facial emotion recognition. *Aggressive Behavior*, *42*(1), 16–28. https://doi.org/10.1002/ab.21602

Dindar, M. (2018). An empirical study on gender, video game play, academic success and complex problem solving skills. *Computers & Education*, *125*, 39–52. https://doi.org/10.1016/j.compedu.2018.05.018

DiStefano, C., & Hess, B. (2005). Using confirmatory factor analysis for construct validation: An empirical review. *Journal of Psychoeducational Assessment*, *23*(3), 225–241. https://doi.org/10.1177%2F073428290502300303

DiStefano, C., & Morgan, G. B. (2014). A comparison of diagonal weighted least squares robust estimation techniques for ordinal data. *Structural Equation Modeling: A Multidisciplinary Journal*, *21*(3), 425–438. https://doi.org/10.1080/10705511.2014.915373

Dodell-Feder, D., Lincoln, S. H., Coulson, J. P., & Hooker, C. I. (2013). Using fiction to assess mental state understanding: A new task for assessing theory of mind in adults. *PLoS One*, *8*(11), Article e81279. https://doi.org/10.1371/journal.pone.0081279

Dodell-Feder, D., & Tamir, D. I. (2018). Fiction reading has a small positive impact on social cognition: A meta-analysis. *Journal of Experimental Psychology: General*, *147*(11), 1713–1727. https://doi.org/10.1037/xge0000395

Donohue, S. E., James, B., Eslick, A. N., & Mitroff, S. R. (2012). Cognitive pitfall! Videogame players are not immune to dual-task costs. *Attention, Perception, & Psychophysics*, *74*(5), 803–809. https://doi.org/10.3758/s13414-012-0323-y

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

Drummond, A., & Sauer, J. D. (2014). Video-games do not negatively impact adolescent academic performance in science, mathematics or reading. *PloS One*, *9*(4), Article e87943. https://doi.org/10.1371/journal.pone.0087943

Drummond, A., & Sauer, J. D. (2015). Daily videogame use and metacognitive knowledge of effective learning strategies. *Psychology of Popular Media Culture*, *4*(4), 342–350. https://doi.org/10.1037/ppm0000049

Drummond, A., & Sauer, J. D. (2019). Timesplitters: Playing video games before (but not after) school on weekdays is associated with poorer adolescent academic performance. A test of competing theoretical accounts. *Computers & Education*, Article 103704. https://doi.org/10.1016/j.compedu.2019.103704

Drummond, A., Sauer, J. D., & Ferguson, C. J. (2020). Do longitudinal studies support long-term relationships between aggressive game play and youth aggressive behaviour? A meta-analytic examination. *Royal Society Open Science*, *7*(7), Article 200373. https://doi.org/10.1098/rsos.200373

Duclos, H., Bejanin, A., Eustache, F., Desgranges, B., & Laisney, M. (2018). Role of context in affective theory of mind in Alzheimer's disease. *Neuropsychologia, 119*, 363–372. https://doi.org/10.1016/j.neuropsychologia.2018.08.025

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

Dunning, D., Griffin, D. W., Milojkovic, J. D., & Ross, L. (1990). The overconfidence effect in social prediction. *Journal of Personality and Social Psychology*, *58*(4), 568–581. https://doi.org/10.1037/0022-3514.58.4.568

Dye, M. W., Green, C. S., & Bavelier, D. (2009). Increasing speed of processing with action video games. *Current Directions in Psychological Science*, *18*(6), 321–326. https://doi.org/10.1111%2Fj.1467-8721.2009.01660.x

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

Edelmann, R. J., & Hampson, S. E. (1979). Changes in non-verbal behaviour during embarrassment. *British Journal of Social and Clinical Psychology*, *18*(4), 385–390. https://doi.org/10.1111/j.2044-8260.1979.tb00908.x

Egleston, B. L., Miller, S. M., & Meropol, N. J. (2011). The impact of misclassification due to survey response fatigue on estimation and identifiability of treatment effects. *Statistics in Medicine, 30*(30), 3560–3572. https://doi.org/10.1002/sim.4377

Engelhardt, C. R., Hilgard, J., & Bartholow, B. D. (2015). Acute exposure to difficult (but not violent) video games dysregulates cognitive control. *Computers in Human Behavior, 45*, 85–92. https://doi.org/10.1016/j.chb.2014.11.089

Entertainment Software Association. (2020). *Essential facts about the video game industry.* Entertainment Software Association.

Eriksson, J. M., Andersen, L. M., & Bejerot, S. (2013). RAADS-14 Screen: Validity of a screening tool for autism spectrum disorder in an adult psychiatric population. *Molecular Autism*, *4*(1), Article 49. https://doi.org/10.1186/2040-2392-4-49

Ewen, J. B., Pillai, A. S., McAuliffe, D., Lakshmanan, B. M., Ament, K., Hallett, M., Crone, N. E., & Mostofsky, S. H. (2016). Practicing Novel, Praxis-Like Movements: Physiological Effects of Repetition. *Frontiers in Human Neuroscience, 10*, Article 22. https://doi.org/10.3389%2Ffnhum.2016.00022

Ewoldsen, D. R., Eno, C. A., Okdie, B. M., Velez, J. A., Guadagno, R. E., & DeCoster, J. (2012). Effect of playing violent video games cooperatively or competitively on subsequent cooperative behavior. *Cyberpsychology, Behavior, and Social Networking*, *15*(5), 277–280. https://doi.org/10.1089/cyber.2011.0308

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

Fekete, J., Pótó, Z., Varga, E., Csulak, T., Zsélyi, O., Tényi, T., & Herold, R. (2020). Persons with schizophrenia misread Hemingway: A new approach to study theory of mind in schizophrenia. *Frontiers in Psychiatry*, *11*, Article 396. https://doi.org/10.3389/fpsyt.2020.00396

Feng, J., Spence, I., & Pratt, J. (2007). Playing an action video game reduces gender differences in spatial cognition. *Psychological Science*, *18*(10), 850–855. https://doi.org/10.1111/j.1467-9280.2007.01990.x

Ferguson, C. J. (2015). Do angry birds make for angry children? A meta-analysis of video game influences on children's and adolescents' aggression, mental health, prosocial behavior, and academic performance. *Perspectives on Psychological Science*, *10*(5), 646–666. https://doi.org/10.1177%2F1745691615592234

Ferguson, C. J., & Beaver, K. M. (2009). Natural born killers: The genetic origins of extreme violence. *Aggression and Violent Behavior*, *14*(5), 286–294. https://doi.org/10.1016/j.avb.2009.03.005

Ferguson, C. J., & Dyck, D. (2012). Paradigm change in aggression research: The time has come to retire the General Aggression Model. *Aggression and Violent Behavior*, *17*(3), 220–228. https://doi.org/10.1016/j.avb.2012.02.007

Ferguson, C. J., & Kilburn, J. (2009). The public health risks of media violence: A meta-analytic review. *The Journal of Pediatrics*, *154*(5), 759–763. https://doi.org/10.1016/j.jpeds.2008.11.033

Ferguson, C. J., Rueda, S. M., Cruz, A. M., Ferguson, D. E., Fritz, S., & Smith, S. M. (2008). Violent video games and aggression: Causal relationship or byproduct of family violence and intrinsic violence motivation? *Criminal Justice and Behavior*, *35*(3), 311–332. https://doi.org/10.1177/0093854807311719

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

Fernández-Abascal, E. G., Cabello, R., Fernández-Berrocal, P., & Baron-Cohen, S. (2013). Test-retest reliability of the 'Reading the Mind in the Eyes' test: A one-year follow-up study. *Molecular Autism*, *4*(1), 1–6. https://doi.org/10.1186/2040-2392-4-33

Finkel, E. J. (2014). The I3 model: Metatheory, theory, and evidence. In *Advances in experimental social psychology* (Vol. 49, pp. 1–104). Elsevier.

Fischer, P., Greitemeyer, T., Morton, T., Kastenmüller, A., Postmes, T., Frey, D., Kubitzki, J., & Odenwälder, J. (2009). The racing-game effect: Why do video racing games increase risk-taking inclinations? *Personality & Social Psychology Bulletin*, *35*(10), 1395–1409. https://doi.org/10.1177/0146167209339628

Fischer, P., Kubitzki, J., Guter, S., & Frey, D. (2007). Virtual driving and risk taking: Do racing games increase risk-taking cognitions, affect, and behaviors? *Journal of Experimental Psychology: Applied*, *13*(1), 22–31. https://doi.org/10.1037/1076-898X.13.1.22

Fitzpatrick, P., Frazier, J. A., Cochran, D., Mitchell, T., Coleman, C., & Schmidt, R. (2018). Relationship between theory of mind, emotion recognition, and social synchrony in

adolescents with and without autism. *Frontiers in Psychology, 9*, 1–13. https://doi.org/10.3389/fpsyg.2018.01337

Flesch, R. (1948). A new readability yardstick. *Journal of Applied Psychology, 32*(3), 221–233. https://doi.org/10.1037/h0057532

Fossati, A., Borroni, S., Dziobek, I., Fonagy, P., & Somma, A. (2018). Thinking about assessment: Further evidence of the validity of the Movie for the Assessment of Social Cognition as a measure of mentalistic abilities. *Psychoanalytic Psychology, 35*(1), 127–141. https://doi.org/10.1037/pap0000130

François, Q., & Rossetti, Y. (2020). What do theory-of-mind tasks actually measure? Theory and practice. *Perspectives on Psychological Science*, *15*(2), 384–396. https://doi.org/10.1177/1745691619896607

Frith, C. D., & Corcoran, R. (1996). Exploring 'theory of mind' in people with schizophrenia. *Psychological Medicine, 26*(3), 521–530. https://doi.org/10.1017/s0033291700035601

Fujita, K., & Itakura. (2008). *Origins of the Social Mind Evolutionary and Developmental Views*. Springer.

Gagnon, D. (1985). Videogames and spatial skills: An exploratory study. *ECTJ*, *33*(4), 263–275. https://doi.org/10.1007/BF02769363

Gallant, C., & Good, D. (2019). Examining the "reading the mind in the eyes test" as an assessment of subtle differences in affective theory of mind after concussion. *The Clinical Neuropsychologist*, 1–22. https://doi.org/10.1080/13854046.2019.1612946

Gallant, C., Lavis, L., & Mahy, C. E. (2020). Developing an understanding of others' emotional states: Relations among affective theory of mind and empathy measures in early childhood. *British Journal of Developmental Psychology*, *38*(2), 151–166. https://doi.org/10.1111/bjdp.12322

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

Gentile, D. A. (2011). The multiple dimensions of video game effects. *Child Development Perspectives*, *5*(2), 75–81. https://doi.org/10.1111/j.1750-8606.2011.00159.x

Gentile, D. A., Anderson, C. A., Yukawa, S., Ihori, N., Saleem, M., Ming, L. K., Shibuya, A., Liau, A. K., Khoo, A., & Bushman, B. J. (2009). The effects of prosocial video games on prosocial behaviors: International evidence from correlational, longitudinal, and experimental studies. *Personality and Social Psychology Bulletin*, *35*(6), 752–763. https://doi.org/10.1177%2F0146167209333045

Geraci, A., Surian, L., Ferraro, M., & Cantagallo, A. (2010). Theory of Mind in patients with ventromedial or dorsolateral prefrontal lesions following traumatic brain injury. *Brain Injury*, *24*(7–8), 978–987. https://doi.org/10.3109/02699052.2010.487477

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%2Farc0000067

Gilbert, F., & Daffern, M. (2011). Illuminating the relationship between personality disorder and violence: Contributions of the General Aggression Model. *Psychology of Violence*, *1*(3), 230–244. https://doi.org/10.1037/a0024089

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), Article 1634326. https://doi.org/10.1080/23311908.2019.1634326

Gliem, J. A., & Gliem, R. R. (2003). *Calculating, interpreting, and reporting Cronbach's alpha reliability coefficient for Likert-type scales*. Midwest Research-to-Practice Conference in Adult, Continuing, and Community Education, Columbus, OH.

Goh, W. D., & Pisoni, D. B. (2003). Effects of lexical competition on immediate memory span for spoken words. *The Quarterly Journal of Experimental Psychology*, *56*(6), 929–954. https://doi.org/10.1080%2F02724980244000710

Goldstein, J., Cajko, L., Oosterbroek, M., Michielsen, M., Van Houten, O., & Salverda, F. (1997). Video games and the elderly. *Social Behavior and Personality: An International Journal*, *25*(4), 345–352. https://doi.org/10.2224/sbp.1997.25.4.345

Goodwin, L. D., & Leech, N. L. (2006). Understanding correlation: Factors that affect the size of r. *The Journal of Experimental Education*, *74*(3), 249–266. https://doi.org/10.3200/JEXE.74.3.249-266

Gopnik, A., & Wellman, H. M. (2012). Reconstructing constructivism: Causal models, Bayesian learning mechanisms, and the theory theory. *Psychological Bulletin*, *138*(6), 1085–1108. https://doi.org/10.1037/a0028044

Gordon, R. M. (1986). Folk psychology as simulation. *Mind & Language*, *1*(2), 158–171. https://doi.org/10.1111/j.1468-0017.1986.tb00324.x

Green, C. S., & Bavelier, D. (2003). Action video game modifies visual selective attention. *Nature*, *423*, 534–537. https://doi.org/10.1038/nature01647

Green, C. S., & Bavelier, D. (2006). Effect of action video games on the spatial distribution of visuospatial attention. *Journal of Experimental Psychology: Human Perception and Performance*, *32*(6), 1465–1478. https://doi.org/10.1037%2F0096-1523.32.6.1465

Green, C. S., & Bavelier, D. (2007). Action-video-game experience alters the spatial resolution of vision. *Psychological Science*, *18*(1), 88–94. https://doi.org/10.1111%2Fj.1467-9280.2007.01853.x

Green, C. S., Sugarman, M. A., Medford, K., Klobusicky, E., & Bavelier, D. (2012). The effect of action video game experience on task-switching. *Computers in Human Behavior*, *28*(3), 984–994. https://doi.org/10.1016/j.chb.2011.12.020

Greenfield, P. M., DeWinstanley, P., Kilpatrick, H., & Kaye, D. (1994). Action video games and informal education: Effects on strategies for dividing visual attention. *Journal of Applied Developmental Psychology*, *15*(1), 105–123. https://doi.org/10.1016/0193-3973(94)90008-6

Greitemeyer, T., & Osswald, S. (2011). Playing prosocial video games increases the accessibility of prosocial thoughts. *The Journal of Social Psychology*, *151*(2), 121–128. https://doi.org/10.1080/00224540903365588

Griffith, J. L., Voloschin, P., Gibb, G. D., & Bailey, J. R. (1983). Differences in eye-hand motor coordination of video-game users and non-users. *Perceptual and Motor Skills*, *57*(1), 155–158. https://doi.org/10.2466/pms.1983.57.1.155

Happé, F. G. (1994). An advanced test of theory of mind: Understanding of story characters' thoughts and feelings by able autistic, mentally handicapped, and normal children and adults. *Journal of autism and Developmental disorders*, *24*(2), 129-154. https://doi.org/10.1007/BF02172093

Harper, J. D., Kaiser, S., Ebrahimi, K., Lamberton, G. R., Hadley, H. R., Ruckle, H. C., & Baldwin, D. D. (2007). Prior video game exposure does not enhance robotic surgical performance. *Journal of Endourology*, *21*(10), 1207–1210. https://doi.org/10.1089/end.2007.9905

Hartanto, A., Toh, W. X., & Yang, H. (2018). Context counts: The different implications of weekday and weekend video gaming for academic performance in mathematics, reading, and science. *Computers & Education, 120,* 51–63. https://doi.org/10.1016/j.compedu.2017.12.007

Hayes, A. F., & Coutts, J. J. (2020). Use omega rather than Cronbach's alpha for estimating reliability. But…. *Communication Methods and Measures*, *14*(1), 1–24.

Hemingway, E. (1927a). A simple enquiry. In *Men without women*. Simon and Schuster.

Hemingway, E. (1927b). Ten indians. In *Men without women*. Simon and Schuster.

Hemingway, E. (1987). *The complete short stories of Ernest Hemingway*. Charles Scribner's Sons.

Hemingway, E. (2003). The end of something. In *In our time*. Scribner.

Henry, J. D., Phillips, L. H., Ruffman, T., & Bailey, P. E. (2013). A meta-analytic review of age differences in theory of mind. *Psychology and Aging*, *28*(3), 826–839. https://doi.org/10.1037/a0030677

Higgins, W. C., Ross, R. M., Langdon, R., & Polito, V. (2022). The "Reading the Mind in the Eyes" Test shows poor psychometric properties in a large, demographically representative US sample. *Assessment*, 1–13. https://doi.org/10.1177/10731911221124342

Holfeld, B., Cicha, R. J., & Ferraro, F. (2015). Executive function and action gaming among college students. *Current Psychology*, *34*(2), 376–388. https://doi.org/10.1007/s12144-014-9263-0

Holmefur, M., Aarts, P., Hoare, B., & Krumlinde-Sundholm, L. (2009). Test-retest and alternate forms reliability of the assisting hand assessment. *Journal of Rehabilitation Medicine, 41*(11), 886–891. https://doi.org/10.2340/16501977-0448

Holt, A., Bounekhla, K., Welch, C., & Polatajko, H. (2022). “Unheard minds, again and again”: Autistic insider perspectives and theory of mind. *Disability and Rehabilitation*, *44*(20), 5887–5897. https://doi.org/10.1080/09638288.2021.1949052

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, *6*(1), 1–55. https://doi.org/10.1080/10705519909540118

Hussey, I., & Hughes, S. (2020). Hidden invalidity among 15 commonly used measures in social and personality psychology. *Advances in Methods and Practices in Psychological Science*, *3*(2), 166–184. https://doi.org/10.1177%2F2515245919882903

Hutchins, T. L., Lewis, L., Prelock, P. A., & Brien, A. (2021). The Development and Preliminary Psychometric Evaluation of the Theory of Mind Inventory: Self Report—Adult (ToMI: SR-Adult). *Journal of Autism and Developmental Disorders*, *51*(6), 1839–1851. https://doi.org/10.1007/s10803-020-04654-6

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

Ioannidis, J. P. (2005). Why most published research findings are false. *PLoS Medicine*, *2*(8), e124.

Jia, R., Steelman, Z. R., & Jia, H. H. (2019). Psychometric assessments of three self-report autism scales (AQ, RBQ-2A, and SQ) for General Adult Populations. *Journal of Autism and Developmental Disorders*, *49*(5), 1949–1965. https://doi.org/10.1007/s10803-019-03880-x

Kember, S. M., & Williams, M. N. (2021). Autism in Aotearoa: Is the RAADS-14 a valid tool for a New Zealand population? *European Journal of Psychological Assessment, 37*(3), 247–257. https://doi.org/10.1027/1015-5759/a000598

Kidd, D. C., & Castano, E. (2013). Reading literary fiction improves theory of mind. *Science, 342*, 377–380. https://doi.org/10.1126/science.1239918

Kim, H.-Y. (2013). Statistical notes for clinical researchers: Assessing normal distribution (2) using skewness and kurtosis. *Restorative Dentistry & Endodontics, 38*(1), 52–54. https://doi.org/10.5395%2Frde.2013.38.1.52

Kincaid, J. P., Fishburne Jr, R. P., Rogers, R. L., & Chissom, B. S. (1975). *Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel*. Naval Education and Training Command.

Kirsh, S. J., & Mounts, J. R. (2007). Violent video game play impacts facial emotion recognition. *Aggressive Behavior, 33*(4), 353–358. http://dx.doi.org/10.1002/ab.20191

Kline, R. B. (2015). *Principles and practice of structural equation modeling* (4th ed.). Guilford Publications.

Kontiola, P., Laaksonen, R., Sulkava, R., & Erkinjuntti, T. (1990). Pattern of language impairment is different in Alzheimer's disease and multi-infarct dementia. *Brain and Language, 38*(3), 364–383. https://doi.org/10.1016/0093-934x(90)90121-v

Koo, T. K., & Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine, 15*(2), 155–163. https://doi.org/10.1016%2Fj.jcm.2016.02.012

Korkmaz, B. (2011). Theory of mind and neurodevelopmental disorders of childhood. *Pediatric Research, 69*(8), 101–108. https://doi.org/10.1203/PDR.0b013e318212c177

Kühberger, A., & Luger-Bazinger, C. (2016). Predicting framed decisions: Simulation or theory? *Psychology, 7*(6), 941–952. http://dx.doi.org/10.4236/psych.2016.76095

Kühn, S., Kugler, D. T., Schmalen, K., Weichenberger, M., Witt, C., & Gallinat, J. (2019). Does playing violent video games cause aggression? A longitudinal intervention study. *Molecular Psychiatry*, *24*(8), 12–20. https://doi.org/10.1038/s41380-018-0031-7

Kulke, L., Johannsen, J., & Rakoczy, H. (2019). Why can some implicit Theory of Mind tasks be replicated and others cannot? A test of mentalizing versus submentalizing accounts. *PloS One*, *14*(3), 1–16. https://doi.org/10.1371/journal.pone.0213772

Kulke, L., Reiß, M., Krist, H., & Rakoczy, H. (2018). How robust are anticipatory looking measures of Theory of Mind? Replication attempts across the life span. *Cognitive Development*, *46*, 97–111. https://doi.org/10.1016/j.cogdev.2017.09.001

Laillier, R., Viard, A., Caillaud, M., Duclos, H., Bejanin, A., de La Sayette, V., Eustache, F., Desgranges, B., & Laisney, M. (2019). Neurocognitive determinants of theory of mind across the adult lifespan. *Brain and Cognition*, *136*, 103588. https://doi.org/10.1016/j.bandc.2019.103588

Larson, A. L., Cycyk, L. M., Carta, J. J., Hammer, C. S., Baralt, M., Uchikoshi, Y., An, Z. G., & Wood, C. (2020). A systematic review of language-focused interventions for young children from culturally and linguistically diverse backgrounds. *Early Childhood Research Quarterly*, *50*, 157–178. https://doi.org/10.1016/j.ecresq.2019.06.001

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

Lee, S., Jacobsen, E. P., Jia, Y., Snitz, B. E., Chang, C.-C. H., & Ganguli, M. (2021). Reading the Mind in the Eyes: A population-based study of social cognition in older

adults. *The American Journal of Geriatric Psychiatry*, *29*(7), 634–642. https://doi.org/10.1016/j.jagp.2020.11.009

Lenhart, A., Kahne, J., Middaugh, E., Macgill, A. R., Evans, C., & Vitak, J. (2008). *Teens, video games, and civics: Teens' gaming experiences are diverse and include significant social interaction and civic engagement*. Pew Internet & American Life Project.

Leys, C., Delacre, M., Mora, Y. L., Lakens, D., & Ley, C. (2019). How to classify, detect, and manage univariate and multivariate outliers, with emphasis on pre-registration. *International Review of Social Psychology*, *32*(1), 1–10. https://doi.org/10.5334/irsp.289

Li, R., Polat, U., Makous, W., & Bavelier, D. (2009). Enhancing the contrast sensitivity function through action video game training. *Nature Neuroscience*, *12*(5), 549–551. https://doi.org/10.1038%2Fnn.2296

Lichtenstein, S., Fischhoff, B., & Phillips, L. D. (1982). Calibration of probabilities: The state of the art to 1980. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases* (pp. 306–334). New York: Cambridge University Press.

Little, R. J. (1988). A test of missing completely at random for multivariate data with missing values. *Journal of the American Statistical Association*, *83*(404), 1198–1202. https://doi.org/10.2307/2290157

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

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

Livingston, L. A., Shah, P., Milner, V., & Happé, F. (2020). Quantifying compensatory strategies in adults with and without diagnosed autism. *Molecular Autism*, *11*(1), 1–10. https://doi.org/10.1186/s13229-019-0308-y

Lynch, J., Aughwane, P., & Hammond, T. M. (2010). Video games and surgical ability: A literature review. *Journal of Surgical Education*, *67*(3), 184–189. https://doi.org/10.1016/j.jsurg.2010.02.010

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

Malmberg, K. J., Lehman, M., Annis, J., Criss, A. H., & Shiffrin, R. M. (2014). Consequences of testing memory. In K. Federmeier (Ed.), *Psychology of learning and motivation* (Vol. 61, pp. 285–313). Elsevier.

Mar, R. A. (2011). The neural bases of social cognition and story comprehension. *Annual Review of Psychology*, *62*(1), 103–134. https://doi.org/10.1146/annurev-psych-120709-145406

Mar, R. A., & Oatley, K. (2008). The function of fiction is the abstraction and simulation of social experience. *Perspectives on Psychological Science*, *3*(3), 173–192. https://doi.org/10.1111%2Fj.1745-6924.2008.00073.x

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

Martin, A. (2009). Semantic Memory. In M. Binder, N. Hirokawa, & U. Windhorst (Eds.), *Encyclopedia of Neuroscience*. Springer.

Martin-Chang, S. L., & Gould, O. N. (2008). Revisiting print exposure: Exploring differential links to vocabulary, comprehension and reading rate. *Journal of Research in Reading*, *31*(3), 273–284. https://doi.org/10.1111/j.1467-9817.2008.00371.x

McGlade, N., Behan, C., Hayden, J., O'Donoghue, T., Peel, R., Haq, F., Gill, M., Corvin, A., O'Callaghan, E., & Donohoe, G. (2008). Mental state decoding v. Mental state reasoning as a mediator between cognitive and social function in psychosis. *The British Journal of Psychiatry*, *193*(1), 77–78. https://doi.org/10.1192%2Fbjp.bp.107.044198

McMaster, K. L., Du, X., & Pétursdóttir, A.-L. (2009). Technical features of curriculum-based measures for beginning writers. *Journal of Learning Disabilities*, *42*(1), 41–60. https://doi.org/10.1177/0022219408326212

Megías-Robles, A., Gutiérrez-Cobo, M. J., Cabello, R., Gómez-Leal, R., Baron-Cohen, S., & Fernández-Berrocal, P. (2020). The 'Reading the mind in the Eyes' test and emotional intelligence. *Royal Society Open Science*, *7*(9), Article 201305. https://doi.org/10.1098/rsos.201305

Mier, D., Lis, S., Neuthe, K., Sauer, C., Esslinger, C., Gallhofer, B., & Kirsch, P. (2010). The involvement of emotion recognition in affective theory of mind. *Psychophysiology*, *47*(6), 1028–1039. https://doi.org/10.1111/j.1469-8986.2010.01031.x

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

Mitchley, N. J., Barber, J., Gray, J. M., Brooks, D. N., & Livingston, M. G. (1998). Comprehension of irony in schizophrenia. *Cognitive Neuropsychiatry, 3*(2), 127–138. https://doi.org/10.1016/j.psychres.2006.04.002

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

Modigliani, A. (1971). Embarrassment, facework, and eye contact: Testing a theory of embarrassment. *Journal of Personality and Social Psychology, 17*(1), 15–24. https://doi.org/10.1037/h0030460

Mol, S. E., & Bus, A. G. (2011). To read or not to read: A meta-analysis of print exposure from infancy to early adulthood. *Psychological Bulletin, 137*(2), 267–296. https://doi.org/10.1037/a0021890

Morris, S. E., & Cuthbert, B. N. (2012). Research Domain Criteria: Cognitive systems, neural circuits, and dimensions of behavior. *Dialogues in Clinical Neuroscience, 14*(1), 29–37. https://doi.org/10.31887%2FDCNS.2012.14.1%2Fsmorris

Müller, N., Baumeister, S., Dziobek, I., Banaschewski, T., & Poustka, L. (2016). Validation of the movie for the assessment of social cognition in adolescents with ASD: Fixation duration and pupil dilation as predictors of performance. *Journal of Autism and Developmental Disorders, 46*(9), 2831–2844. https://doi.org/10.1007/s10803-016-2828-z

Mumper, M. L., & Gerrig, R. J. (2017). Leisure reading and social cognition: A meta-analysis. *Psychology of Aesthetics, Creativity, and the Arts, 11*(1), 109–120. https://doi.org/10.1037/aca0000089

Murphy, K., & Spencer, A. (2009). Playing video games does not make for better visual attention skills. *Journal of Articles in Support of the Null Hypothesis, 6*(1), 1–19.

Navarro Garcia, E. (2021). *Theory of mind measurements and mechanisms: An investigation of construct validity and cognitive processes in theory of mind tasks*. Claremont Graduate University.

NewZoo. (2019). *Global Games Market Report*. NewZoo.

Njomboro, P., Deb, S., & Humphreys, G. W. (2008). Dissociation between decoding and reasoning about mental states in patients with theory of mind reasoning impairments. *Journal of Cognitive Neuroscience, 20*(9), 1557–1564. https://doi.org/10.1162/jocn.2008.20118

Nunnally, J. C. (1994). *Psychometric Theory* (3rd ed.). Tata McGraw-Hill Education.

Oakley, B. F., Brewer, R., Bird, G., & Catmur, C. (2016). Theory of mind is not theory of emotion: A cautionary note on the Reading the Mind in the Eyes Test. *Journal of Abnormal Psychology*, *125*(6), 818–823. https://doi.org/10.1037%2Fabn0000182

Oatley, K. (2016). Fiction: Simulation of social worlds. *Trends in Cognitive Sciences, 20*(8), 618–628. https://doi.org/10.1016/j.tics.2016.06.002

Ocal, T., & Ehri, L. (2017). Spelling ability in college students predicted by decoding, print exposure, and vocabulary. *Journal of College Reading and Learning, 47*(1), 58–74. http://dx.doi.org/10.1080/10790195.2016.1219242

Oei, A. C., & Patterson, M. D. (2014). Playing a puzzle video game with changing requirements improves executive functions. *Computers in Human Behavior*, *37*, 216–228. http://dx.doi.org/10.1016/j.chb.2014.04.046

Oi, M., Tanaka, S., & Ohoka, H. (2013). The relationship between comprehension of figurative language by Japanese children with high functioning autism spectrum disorders and college freshmen's assessment of its conventionality of usage. *Autism Research and Treatment, 2013*, 1–7. http://dx.doi.org/10.1155/2013/480635

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,* Article 1503. https://doi.org/10.3389%2Ffpsyg.2015.01503

Onishi, K. H., & Baillargeon, R. (2005). Do 15-month-old infants understand false beliefs? *Science, 308,* 255–258. https://doi.org/10.1126%2Fscience.1107621

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science, 349,* Article aac4716. https://doi.org/10.1126/science.aac4716

Ou, Y., McGlone, E. R., Camm, C. F., & Khan, O. A. (2013). Does playing video games improve laparoscopic skills? *International Journal of Surgery*, *11*(5), 365–369. http://dx.doi.org/10.1016/j.ijsu.2013.02.020

Panero, M. E., Weisberg, D. S., Black, J., Goldstein, T. R., Barnes, J. L., Brownell, H., & Winner, E. (2016). Does reading a single passage of literary fiction really improve theory of mind? An attempt at replication. *Journal of Personality and Social Psychology*, *111*(5), 46–54. https://doi.org/10.1037/pspa0000064

Park, D. C., Lodi-Smith, J., Drew, L., Haber, S., Hebrank, A., Bischof, G. N., & Aamodt, W. (2014). The impact of sustained engagement on cognitive function in older adults: The Synapse Project. *Psychological Science*, *25*(1), 103–112. https://doi.org/10.1177%2F0956797613499592

Pashler, H., & Wagenmakers, E. (2012). Editors' introduction to the special section on replicability in psychological science: A crisis of confidence? *Perspectives on Psychological Science*, *7*(6), 528–530. https://doi.org/10.1177/1745691612465253

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/10.1016/j.cognition.2006.04.012

Pichon, S., Bediou, B., Antico, L., Jack, R., Garrod, O., Sims, C., Green, C. S., Schyns, P., & Bavelier, D. (2018). Emotion perception in habitual players of action video games. *Emotion*, *21*(6), 1324–1339. https://doi.org/10.1037/emo0000740

Pisula, E., & Ziegart-Sadowska, K. (2015). Social communication and language deficits in parents and siblings of children with ASD—A short review. In *Autism Spectrum Disorder—Recent Advances*. Intech Open.

Poletti, M., & Adenzato, M. (2013). Theory of mind in non-autistic psychiatric disorders of childhood and adolescence. *Clinical Neuropsychiatry*, *10*(5), 188–196.

Popper, K. (1983). *Realism and the aim of science: From the postscript to the logic of scientific discovery*. Routledge.

Poulin-Dubois, D., & Yott, J. (2018). Probing the depth of infants' theory of mind: Disunity in performance across paradigms. *Developmental Science*, *21*(4), 1–11. https://doi.org/10.1111/desc.12600

Powers, K. L., Brooks, P. J., Aldrich, N. J., Palladino, M. A., & Alfieri, L. (2013). Effects of video-game play on information processing: A meta-analytic investigation. *Psychonomic Bulletin & Review*, *20*(6), 1055–1079. https://doi.org/10.3758/s13423-013-0418-z

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

Quan, F., Yang, R., & Xia, L.-X. (2021). The longitudinal relationships among agreeableness, anger rumination, and aggression. *Current Psychology*, *40*(1), 9–20. https://doi.org/10.1007/s12144-020-01030-6

Richman, M. J., & Unoka, Z. (2015). Mental state decoding impairment in major depression and borderline personality disorder: Meta-analysis. *The British Journal of Psychiatry*, *207*(6), 483–489. https://doi.org/10.1192/bjp.bp.114.152108

Rogers, K., Dziobek, I., Hassenstab, J., Wolf, O. T., & Convit, A. (2007). Who Cares? Revisiting empathy in Asperger Syndrome. *Journal of Autism and Developmental Disorders*, *37*(4), 709–715. https://doi.org/10.1007/s10803-006-0197-8

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

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

Sabbagh, M. A. (2004). Understanding orbitofrontal contributions to theory-of-mind reasoning: Implications for autism. *Brain and Cognition*, *55*(1), 209–219. https://doi.org/10.1016/j.bandc.2003.04.002

Sachdev, P. S., Blacker, D., Blazer, D. G., Ganguli, M., Jeste, D. V., Paulsen, J. S., & Petersen, R. C. (2014). Classifying neurocognitive disorders: The DSM-5 approach. *Nature Reviews Neurology*, *10*(11), 634–642. https://doi.org/10.1038/nrneurol.2014.181

Salkind, N. J. (2010). *Encyclopedia of research design* (Vol. 1). Sage Publications.

Salthouse, T. A. (2000). Aging and measures of processing speed. *Biological Psychology*, *54*(1–3), 35–54. https://doi.org/10.1016/s0301-0511(00)00052-1

Samur, D., Tops, M., & Koole, S. L. (2018). Does a single session of reading literary fiction prime enhanced mentalising performance? Four replication experiments of Kidd and Castano (2013). *Cognition and Emotion*, *32*(1), 130–144. https://doi.org/10.1080/02699931.2017.1279591

San José Cáceres, A., Keren, N., Booth, R., & Happé, F. (2014). Assessing Theory of Mind Nonverbally in Those With Intellectual Disability and ASD: The P enny H iding G ame. *Autism Research*, *7*(5), 608–616. https://doi.org/10.1002/aur.1405

Sarmet, M. M., & Pilati, R. (2016). The effect of digital games on behavior: Analysis of the General Learning Model. *Trends in Psychology*, *24*(1), 17–31. http://dx.doi.org/10.9788/TP2016.1-03

Satyen, L. (2005). *Video game playing: Its effects on divided attention, encoding and retrieval processes of human memory* [Doctoral dissertation]. Victoria University.

Savalei, V., & Rhemtulla, M. (2013). The performance of robust test statistics with categorical data. *British Journal of Mathematical and Statistical Psychology*, *66*(2), 201–223. https://doi.org/10.1111/j.2044-8317.2012.02049.x

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

Scharrer, E., Kamau, G., Warren, S., & Zhang, C. (2018). Violent video games do contribute to aggression. In C. J. Ferguson (Ed.), *Video game influences on aggression, cognition, and attention* (pp. 5–21). Springer.

Schedler, C. (2013). *Border Modernism*. Routledge.

Scheffer, J. (2002). Dealing with missing data. *Research Letters in the Information and Mathematical Sciences*, *3*, 153–160.

Schenk, S., Lech, R. K., & Suchan, B. (2017). Games people play: How video games improve probabilistic learning. *Behavioural Brain Research, 335*, 208–214. https://doi.org/10.1016/j.bbr.2017.08.027

Schneider, D., Slaughter, V. P., & Dux, P. E. (2017). Current evidence for automatic Theory of Mind processing in adults. *Cognition, 162*, 27–31. https://doi.org/10.1016/j.cognition.2017.01.018

Schooler, J. W. (2014). Metascience could rescue the 'replication crisis.' *Nature, 515*, 9–9. https://doi.org/10.1038/515009a

Schurz, M., & Perner, J. (2015). An evaluation of neurocognitive models of theory of mind. *Frontiers in Psychology, 6*, 1–9. https://doi.org/10.3389/fpsyg.2015.01610

Schuwerk, T., Vuori, M., & Sodian, B. (2015). Implicit and explicit theory of mind reasoning in autism spectrum disorders: The impact of experience. *Autism*, *19*(4), 459–468. https://doi.org/10.1177/1362361314526004

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

Seçer, I., & Satyen, L. (2013). Training skills of divided attention among older adults. *Journal of Articles in Support of the Null Hypothesis*, *9*(2), 61–78.

Shamay-Tsoory, S., & 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

Shamay-Tsoory, S., Harari, H., Aharon-Peretz, J., & Levkovitz, Y. (2010). The role of the orbitofrontal cortex in affective theory of mind deficits in criminal offenders with

psychopathic tendencies. *Cortex*, *46*(5), 668–677. https://doi.org/10.1016/j.cortex.2009.04.008

Shamay-Tsoory, S., Tomer, R., Berger, B., & Aharon-Peretz, J. (2003). Characterization of empathy deficits following prefrontal brain damage: The role of the right ventromedial prefrontal cortex. *Journal of Cognitive Neuroscience*, *15*(3), 324–337. https://doi.org/10.1162/089892903321593063

Shamay-Tsoory, S., Tomer, R., Berger, B. D., Goldsher, D., & Aharon-Peretz, J. (2005). Impaired "affective theory of mind" is associated with right ventromedial prefrontal damage. *Cognitive and Behavioral Neurology*, *18*(1), 55–67. https://doi.org/10.1097/01.wnn.0000152228.90129.99

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

Shao, Z., Janse, E., Visser, K., & Meyer, A. (2014). What do verbal fluency tasks measure? Predictors of verbal fluency performance in older adults. *Frontiers in Psychology*, *5*, 1–10. https://doi.org/10.3389/fpsyg.2014.00772

Sherman, G. D., Lerner, J. S., Renshon, J., Ma-Kellams, C., & Joel, S. (2015). Perceiving others' feelings: The importance of personality and social structure. *Social Psychological and Personality Science*, *6*(5), 559–569. https://doi.org/10.1177/1948550614567358

Simas, R., Maestri, F., & Normando, D. (2014). Controlling false positive rates in research and its clinical implications. *Dental Press Journal of Orthodontics*, *19*(3), 24–25. https://doi.org/10.1590%2F2176-9451.19.3.024-025.ebo

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2016). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, *22*(11), 1359–1366. https://doi.org/10.1037/14805-033

Smogorzewska, J., Szumski, G., & Grygiel, P. (2018). Same or different? Theory of mind among children with and without disabilities. *PLoS One*, *13*(10), Article e0202553. https://doi.org/10.1371/journal.pone.0202553

Southgate, V., Senju, A., & Csibra, G. (2007). Action anticipation through attribution of false belief by 2-year-olds. *Psychological Science*, *18*(7), 587–592. https://doi.org/10.1111/j.1467-9280.2007.01944.x

Spaulding, S. (2012). Mirror neurons are not evidence for the Simulation Theory. *Synthese*, *189*(3), 515–534. https://doi.org/10.1007/s11229-012-0086-y

Stanovich, K. E., & West, R. F. (1989). Exposure to print and orthographic processing. *Reading Research Quarterly*, 402–433.

Stevenson, J. L., & Hart, K. R. (2017). Psychometric properties of the autism-spectrum quotient for assessing low and high levels of autistic traits in college students. *Journal of Autism and Developmental Disorders*, *47*(6), 1838–1853. https://doi.org/10.1007/s10803-017-3109-1

Strobach, T., Frensch, P. A., & Schubert, T. (2012). Video game practice optimizes executive control skills in dual-task and task switching situations. *Acta Psychologica*, *140*(1), 13–24. https://doi.org/10.1016/j.actpsy.2012.02.001

Summers, J., Cheng, H.-Y., Lin, H.-H., Barnard, L. T., Kvalsvig, A., Wilson, N., & Baker, M. G. (2020). Potential lessons from the Taiwan and New Zealand health responses to the COVID-19 pandemic. *The Lancet Regional Health-Western Pacific*, *4*, Article 100044. https://doi.org/10.1016/j.lanwpc.2020.100044

Sze, J. A., Goodkind, M. S., Gyurak, A., & Levenson, R. W. (2012). Aging and emotion recognition: Not just a losing matter. *Psychology and Aging*, *27*(4), 940–950. https://doi.org/10.1037%2Fa0029367

Tabachnick, B. G., Fidell, L. S., & Ullman, J. B. (2007). *Using multivariate statistics* (Vol. 5). Pearson.

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

Trisolini, D. C., Petilli, M. A., & Daini, R. (2018). Is action video gaming related to sustained attention of adolescents? *Quarterly Journal of Experimental Psychology*, *71*(5), 1033–1039. http://dx.doi.org/10.1080/17470218.2017.1310912

Trott, S., & Bergen, B. (2020). When do comprehenders mentalize for pragmatic inference? *Discourse Processes*, *57*(10), 900–920. https://doi.org/10.1080/0163853X.2020.1822709

Turner, R., & Felisberti, F. M. (2017). Measuring mindreading: A review of behavioral approaches to testing cognitive and affective mental state attribution in neurologically typical adults. *Frontiers in Psychology*, *8*, 1–7. https://doi.org/10.3389/fpsyg.2017.00047

Unsworth, N., Redick, T. S., McMillan, B. D., Hambrick, D. Z., Kane, M. J., & Engle, R. W. (2015). Is playing video games related to cognitive abilities? *Psychological Science*, *26*(6), 759–774. https://doi.org/10.1177/0956797615570367

van Ravenzwaaij, D., Boekel, W., Forstmann, B. U., Ratcliff, R., & Wagenmakers, E.-J. (2014). Action video games do not improve the speed of information processing in simple perceptual tasks. *Journal of Experimental Psychology: General*, *143*(5), 1–24. http://dx.doi.org/10.1037/a0036923

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

Völter, C. J., & Call, J. (2014). Younger apes and human children plan their moves in a maze task. *Cognition*, *130*(2), 186–203. https://doi.org/10.1016/j.cognition.2013.10.007

Wang, P., Liu, H.-H., Zhu, X.-T., Meng, T., Li, H.-J., & Zuo, X.-N. (2016). Action video game training for healthy adults: A meta-analytic study. *Frontiers in Psychology*, *7*, 1–13. https://doi.org/10.3389%2Ffpsyg.2016.00907

Waris, O., Jaeggi, S. M., Seitz, A. R., Lehtonen, M., Soveri, A., Lukasik, K. M., Söderström, U., Hoffing, R. C., & Laine, M. (2019). Video gaming and working memory: A large-scale cross-sectional correlative study. *Computers in Human Behavior*, *97*, 94–103. https://doi.org/10.1016/j.chb.2019.03.005

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

Weber, N., & Brewer, N. (2004). Confidence-accuracy calibration in absolute and relative face recognition judgments. *Journal of Experimental Psychology: Applied*, *10*(3), 156–172. https://doi.org/10.1037/1076-898X.10.3.156

West, G. L., Stevens, S. A., Pun, C., & Pratt, J. (2008). Visuospatial experience modulates attentional capture: Evidence from action video game players. *Journal of Vision*, *8*(16), 1–9. https://doi.org/10.1167/8.16.13

West, R. F., Stanovich, K. E., & Mitchell, H. R. (1993). Reading in the real world and its correlates. *Reading Research Quarterly*, *28*(1), 35–50. https://doi.org/10.2307/747815

Wolfe, J., Kar, K., Perry, A., Reynolds, C., Gradisar, M., & Short, M. A. (2014). Single night video-game use leads to sleep loss and attention deficits in older adolescents. *Journal of Adolescence, 37*(7), 1003–1009. https://doi.org/10.1016/j.adolescence.2014.07.013

Woolley, J. D., & Van Reet, J. (2006). Effects of context on judgments concerning the reality status of novel entities. *Child Development, 77*(6), 1778–1793. https://doi.org/10.1111/j.1467-8624.2006.00973.x

Wu, S., & Spence, I. (2013). Playing shooter and driving videogames improves top-down guidance in visual search. *Attention, Perception, & Psychophysics*, *75*(4), 673–686. https://doi.org/10.3758/s13414-013-0440-2

Xia, Y., & Yang, Y. (2019). RMSEA, CFI, and TLI in structural equation modeling with ordered categorical data: The story they tell depends on the estimation methods. *Behavior Research Methods*, *51*(1), 409–428. https://doi.org/10.3758/s13428-018-1055-2

Yates, B. T., & Taub, J. (2003). Assessing the costs, benefits, cost-effectiveness, and cost-benefit of psychological assessment: We should, we can, and here's how. *Psychological Assessment*, *15*(4), 478–495. http://dx.doi.org/10.1037/1040-3590.15.4.478

Zillmann, D. (1988). Cognition-excitation interdependences in aggressive behavior. *Aggressive Behavior*, *14*(1), 51–64. https://doi.org/10.1002/1098-2337(1988)14:1%3C51::AID-AB2480140107%3E3.0.CO;2-C

APPENDIX A

Calculation of the Flesch Reading Ease Score and Flesch-Kincaid Grade Level

Test	Formula
FRES	206.835 - 1.015 × (Total Words / Total Sentences) - 84.6 × (Total Syllables / Total Words)
FKGL	0.39 × (Total Words / Total Sentences) + 11.8 × (Total Syllables / Total Words) − 15.59

Note. FRES = Flesch Reading Ease Score; FKGL = Flesch-Kincaid Grade Level

APPENDIX B-1

Short Story Task Questions and Marking Rubric

Short Story Task Scoring Questions, Instructions, and Rubric

General Instructions:

- If the participant provides a response that is patently wrong, i.e., a "spoiled" response, unless they take it back or qualify their response in a way that would give them a score of 1 or 2 (i.e., back-track on their spoiled response), it should be given a score of 0.
- If the participant provides a response that would be scored a 1 and a 2, score it as a 2.
- If the participant provides a response that is accurate, but does not answer the question, it should be scored a 0.

1) *COMPREHENSION*: What do Nick and Marjorie observe on the shoreline as they are rowing to the point to set their fishing lines?

2 – An abandoned (lumber) mill; the white limestone foundations of the (lumber) mill; a broken down/old (lumber) mill; an abandoned (lumbering) town; Horton's Bay

1 – Description of shoreline; swampy growth/swampy meadow; (lumber) mill; AND no mention of town or buildings/structures

0 – None of the responses in 2 or 1

2) *COMPREHENSION*: What does Nick mean when he says, "They aren't striking?"

2 – The fish aren't eating the bait; the fish aren't feeding; fishing isn't going well

1 – Any response that mentions fishing, but does not speak to the fact that the fish aren't taking the bait

0 – Any response that doesn't convey that the fish aren't taking the bait or that fishing is not going well or that doesn't mention fishing at all

3) *COMPREHENSION*: Nick and Marjorie have a pail of perch for what purpose?

2 – Bait; catching fish

1 – Any response that conveys understanding of its use for some aspect of fishing without being explicit about its function as bait for catching fish

0 – Any response that doesn't convey understanding of its use for fishing

4) *COMPREHENSION*: Do Marjorie's actions suggest that she is experienced or inexperienced at fishing? What makes you say that?

2 – Experienced/somewhat experienced/somewhat inexperienced (depending on whether they interpret her skills relative to Nick versus most people) with following possible justifications: she says she loves fishing, can prepare the bait, holds the line in her mouth, knows how to steer the boat, knows where to cast the line; she can prepare the bait, although perfectly or as good as Nick, often asks Nick if she's doing things correctly and Nick often corrects her; only somewhat experienced because she needs to constantly ask Nick and doesn't do things as well as Nick

1 – Somewhat experienced/inexperienced, but with poor justification (i.e., none of the justifications mentioned above); inexperienced with justifications mentioned above

0 – Inexperienced without qualifying response; if response is "inexperienced," but then back tracks after giving justification and changes to somewhat experienced/inexperienced, should be counted as **1** or **2** depending on justification

5) *THEORY OF MIND*: Why does Nick say to Marjorie, "You know everything"?

2 – He's being sarcastic/cynical/intentionally mean AND wants to get Marjorie upset/sad/mad/annoyed; provoke a fight or provoke Marjorie so that she breaks up with him so he can blame the breakup on her

1 – He's unhappy with the relationship; wants to end the relationship; He's annoyed/nervous about the situation/impending breakup; he's being sarcastic/cynical (no mention of consequences, i.e., what Marjorie's reaction will be)

0 – He thinks Marjorie is a know-it-all; He's just being mean; He's a mean person

6) *THEORY OF MIND*: Why does Marjorie reply, "Oh Nick, please cut it out! Please, please don't be that way!"?

2 – She knows Nick is trying to provoke a fight/intentionally giving her a hard time and doesn't want to have a confrontation; she senses that Nick might break-up with her

1 – They have had this type of conversation before; she doesn't want to fight; she doesn't want to ruin a nice day

0 – None of the responses in 2 or 1

7) *THEORY OF MIND*: Why is Nick afraid to look at Marjorie?

2 – *response needs to reference Marjorie's possible reaction to what he's saying;* He knows she is hurt/upset by his comment, and he is afraid of her reaction/doesn't want to see the hurt in her face; is afraid of her judgment of him

1 – *some response that conveys his feelings without referencing how Marjorie's reactions affect his feelings;* he is uncomfortable with the way the conversation is heading; he feels

guilty/shameful/sad; he's about to break up with her and it's easier not to look at her; he's afraid he's making the wrong decision by breaking up with her

0 – he doesn't want Marjorie to see *his* reaction; none of the responses in 2 or 1

8) *THEORY OF MIND*: What does Nick mean when he says, "It isn't fun anymore"?

2 – He's tired of the relationship; he wants to end the relationship; their relationship/love isn't fun anymore; the relationship is no longer enjoyable/bringing him happiness (response can mention fishing as example of how nothing they do together is fun any longer; response can also be "Could be the relationship, or could be fishing"); being with her (as long as "being with" refers to being with her more globally, like in the context of the relationship, and not on this specific fishing trip or being with her in this very moment)

1 – A response that only partially captures the understanding that he's dissatisfied with

their relationship or alludes to the relationship without explicitly acknowledging dissatisfaction with the relationship per se, e.g. being around her/spending time with her

0 – going fishing with her ONLY; None of the responses in 2 or 1

9) *THEORY OF MIND*: Why does Marjorie sit with her back toward Nick when she asks, "Isn't love any fun?"

2 – She knows Nick is about to end their relationship; she is afraid of his answer because she knows it's going to be bad/hurtful/not what she wants to hear; she's trying to protect herself from his response because she knows it's bad; she's afraid of showing him how vulnerable/hurt/upset she is

1 – She's upset/mad/afraid of crying; she is uncomfortable with the conversation

0 – None of the responses in 2 or 1

10) *<u>THEORY OF MIND</u>:* Why does Marjorie take the boat and leave (*1 point*) and what is she feeling at that moment (*1 point*)?

2 – She realizes her relationship with Nick is over, she wants space to herself, she doesn't want Nick to see her upset/vulnerable AND she's feeling upset/sad/angry/disappointed/rejected (negative affect)

1 – Any response that doesn't fully and accurately convey Marjorie's understanding that the relationship is over AND feels anger/disappointment/sadness etc. (negative affect) – i.e., the answer needs to convey understanding of Marjorie's negative affect, but does not convey Marjorie's understanding that the relationship is over or vice-versa

0 – None of the responses in 2 or 1

11) *<u>THEORY OF MIND</u>:* Who is Bill and what does he reveal when he asks Nick, "Did she go alright? … Have a scene?"?

2 – Bill is a friend/lover of Nick; Bill knew that Nick was going to break up with Marjorie and would likely be upset/angry/fight with Nick in response

1 – Any response that misidentifies Bill's relation to Nick (or just says Bill has some form of relationship with Nick/Bill knows Nick) OR doesn't acknowledge that Bill knew something in advance

0 – None of the responses in 2 or 1

12) *<u>THEORY OF MIND</u>:* What is Nick feeling when he says, "Oh, go away, Bill! Go away for a while"?

2 – He feels guilty/sad/upset about hurting Marjorie (negative affect in relation to the fact that he just broke up with Marjorie) and needs his space to process things/doesn't want to talk about it with Bill

1 – Any response that describes that Nick is experiencing negative affect, but doesn't put the affect in the context of him breaking up with/hurting/upsetting Marjorie (answer can include he wants to be left alone), e.g. he's upset/sad/angry and wants to be left alone

0 – ONLY he wants to be left alone/he wants some space to think about things/process things; none of the responses in 2 or 1

13) *COMPREHENSION*: The story is called "The End of Something." What is the title referring to?

2 – the end of Nick and Marjorie's relationship; the end of innocence or happiness; the end of being able to blame someone else for your own actions/decisions; the end of one of these things and (but doesn't have to include) the end of Horton's Bay as a bustling lumbering town/the end of the mill

1 – mentions ONLY the end of something related to Horton's Bay/the mill

0 – None of the responses in 2 or 1

APPENDIX B-2

Short Story Task-B Questions and Marking Rubric (Original)

Short Story Task B Scoring Questions, Instructions, and Rubric

General Instructions:

- If the participant provides a response that is patently wrong, i.e., a "spoiled" response, unless they take it back or qualify their response in a way that would give them a score of 1 or 2 (i.e., back-track on their spoiled response), it should be given a score of 0.
- If the participant provides a response that would be scored a 1 and a 2, score it as a 2.
- If the participant provides a response that is accurate, but does not answer the question, it should be scored a 0.

1) ***COMPREHENSION:*** **What can be observed outside the hut's windows?**

   2 – Snow; snow that is higher than the window; melted snow; snow melted into a trench

   1 – Sunlight; light; a trench; AND no mention of snow

   0 – Neither 1 nor 2.

2) ***COMPREHENSION:*** **The major has a saucer of oil for what purpose?**

   2 – To sooth his burns/skin; to moisten his skin/burns; to ease the pain of his burns

   1 – Reference to its use for his skin/burns without stating it is to moisten/sooth/ease pain (e.g., for his skin/burns)

   0 – Neither 1 nor 2

3) ***COMPREHENSION:*** **Do the adjutant's actions suggest he is hardworking or lazy? What makes you think that?**

2 – Hardworking/somewhat hardworking/somewhat lazy. Justifications: he does his work before relaxing; he cannot relax/read/smoke without first completing his work; he begins to stop working, but then stops procrastinating/going of task before he finishes his work

1 – Hardworking/somewhat hardworking/somewhat lazy/ but with poor justification (i.e., none of those mentioned above); lazy with justifications mentioned above.

0 – Lazy without giving any of the justifications outlined in 2.

4) ***COMPREHENSION:*** **What does the adjutant mean when he says, "Be soft, Pinin … The major is sleeping"?**

2 – Pinin needs to be quiet/not make any noise so as to not wake the major; Pinin needs to move slowly/work slowly/work quietly to not wake the major

1 – Any reference to Pinin needing to be quiet/work slow without referencing that it is to not wake the major.

0 – Neither 1 nor 2.

5) ***THEORY OF MIND:*** **What does Pinin mean when he replies, "I have been with girls."**

2 – Pinin is attempting to deflect the major's questions/he is being purposefully ambiguous BECAUSE they make him uncomfortable/he does not want to answer such a personal question/he is trying to hide something.

1 – Pinin is trying to deflect the major's question (no reason as to why he may be doing this is given)

0 – The statement is literally interpreted; neither 1 nor 2.

6) ***THEORY OF MIND:*** **Why does the major say "Tonani … can you hear me talking?"**

2 – To protect the privacy of himself and/or Pinin AND to put Pinin at ease/make him more comfortable to talk openly by showing him Tonani cannot hear them/to reassure both of them Tonani cannot hear them.

1 –Makes mention of the action intention (i.e., to make Pinin talk more openly/protect the majors/Pinin's privacy) without acknowledging emotions (e.g. put them at ease/make Pinin less nervous etc).

0 – Repetition of quote explicit meaning "to see if Tonani could hear them." without directly referencing underlying intentions or characters emotions; neither 1 nor 2.

7) ***THEORY OF MIND:*** **Why does the major say "All right … You needn't be superior."?**

2 – The major believes that Pinin is suggesting that homosexuality is inferior/he is above homosexuality; the major is acting defensively as he perceives Pinin to believe he is above him/morally superior for not being homosexual; the major believes Pinin is brazenly asserting his awareness of the major's sexuality and that Pinin believes is above it; the major thinks Pinin is suggesting heterosexuals are above homosexuals.

1 – Any answer that does not accurately interpret the major's perception of Pinin's intention; failure to acknowledge that it is the major's perception of Pinin's intention (e.g., Pinin believed that homosexuality was inferior).

0 – Neither 1 nor 2.

8) ***THEORY OF MIND:*** **Why does Pinin look at the floor when the major asks him "And you really don't want-" and "That your great desire isn't really-"?**

2 – Pinin is attempting to conceal his true feelings from the major by avoiding his questions through both lack of verbal and/or hiding behavioural response(s). This may be to either hide his thoughts/disdain of the major's homosexuality or hide his own sexual orientation/he is uncomfortable with the conversation AND is trying to hide this from the major.

1 – The conversation makes him uncomfortable (no mention of hiding responses from the major)

0 – Neither 1 nor 2.

9) *THEORY OF MIND:* Why is the major 'really relieved'?

2 – He does not have to worry about him and Pinin concealing a homosexual relationship from others/in the military; the major was not overly interested in relationship with Pinin; and he is relieved for the conversation to be over.

1 – Acknowledgement of relief without reference to worry/disinterest/etc in a relationship with Pinin (e.g., only responding "he is relieved for the conversation to be over.")

0 – Life in the military was too complicated; Neither 1 nor 2.

10) *THEORY OF MIND:* What is Pinin feeling when he leaves the major's room and walks outside

2 – Pinin is feeling embarrassed/ashamed/upset/discomfort for being propositioned or thought homosexual by the major/to hide his true feelings; Pinin is naïve and is shocked at being propositioned so openly/learning that his superior is a homosexual

1 – Correct reference to Pinin's feelings (or a related synonym to those above) without putting the feeling into context

0 – No mention of emotion, incorrect emotion

11) *THEORY OF MIND:* Who is James and what does he reveal when he asks, "So the major propositioned you too?"

2 – James is a friend/lover/colleague of Pinin's; James is implying the Pinin is not the first person the major has made a sexual proposition to/questioned their sexuality.

1 – Any response that misidentifies James' relationship to Pinin OR doesn't acknowledge that Pinin is not the first person the major has propositioned.

0 – Neither 1 nor 2.

12) *THEORY OF MIND:* Why does Pinin avoid eye contact with James?

2 – He feels shame/embarrassment/etc. from his interaction with the major AND does not want James to see this/does not want to see whether James is judging him/is afraid to see if James is judging him

1 – A mention of the discomfort/shame/embarrassment linked to eye contact WITHOUT making mention of him hiding this from James/mention of James

0 – Neither 1 nor 2.

13) *COMPREHENSION:* The story is called "A Simple Enquiry." What is the title referring to?

2 – The major's sexual proposition towards Pinin; the major enquiring about a potential relationship with Pinin; the major's enquiring about Pinin's sexuality/love interest/sexual history.

1 – Reference to the conversation between the major and Pinin without explicit mention of the conversations purpose.

0 – Neither 1 nor 2.

APPENDIX C-1

Information Sheet for Study 1

MASSEY UNIVERSITY
TE KUNENGA KI PUREHUROA
UNIVERSITY OF NEW ZEALAND

The Pilot of the Short Story Task B

INFORMATION SHEET

Description:

In this study, you will complete a survey that will ask a broad range of questions about your understanding of pieces of literary fiction, video game use, and personal traits. Your participation should take no more than 55 minutes. Your participation is voluntary and you have the right to refuse to answer any question or withdraw from the study at any time without penalty. After you complete the survey you will be given an explanation of the study and provided with monetary compensation. You may also provide an email address which we will use to let you know about the findings of this research.

Participants should: Be aged 16 or above. Not have a current neurological, psychiatric, or developmental disorder diagnosis. Not have severe visual impairment or epilepsy. Play video games for a minimum of one hour a week.

This study consists of an online questionnaire. The questionnaire consists of a number of multiple-choice and open-answer questions. Questions will be presented in a number of sections. If you choose to participate, please complete all sections in one sitting. You will not be able to resume at another time from where you left off. Your answers will not be saved until you complete all sections of the questionnaire.

Once the data have been analysed, we will ensure that we remove from the data set any information that might inadvertently include any identifying information. We will then make this non-identifiable data available to other researchers and might post it to an online repository.

If you wish to participate in this study and all of your questions have been answered, then please move to the next screen. If you do not wish to participate in this study, please return your submission on Prolific by selecting the 'Stop without completing' button.

Contact Information

Researcher
Joshua Robinson
Doctor of Clinical Psychology Student
Massey University
Palmerston North
New Zealand
Email: JoshuaResearch123@gmail.com

Supervisor
Dr Michael Philipp
School of Psychology
Massey University
Palmerston North
New Zealand
+64 6 951-8086
Email: m.philipp@massey.ac.nz

Supervisor
Dr Aaron Drummond
School of Psychology
Massey University
Palmerston North
New Zealand
+64 6 356-9099
Email: a.drummond@massey.ac.nz

Supervisor
Dr Clifford Van Ommen
School of Psychology
Massey University
Auckland
New Zealand
+64 6 951-0800
Email: c.vanommen@massey.ac.nz

Massey University School of Psychology – Te Kura Hinengaro Tangata
Palmerston North, New Zealand
T +64 6 3569-099 ext 85071 : W psychology.massey.ac.nz

This project has been evaluated by peer review and judged to be low risk. Consequently it has not been reviewed by one of the University's Human Ethics Committees. The researcher named in this document is responsible for the ethical conduct of this research.
If you have any concerns about the conduct of this research that you want to raise with someone other than the researcher(s), please contact Professor Craig Johnson, Director (Research Ethics), email humanethics@massey.ac.nz.

Next

APPENDIX C-2

Information Sheet for Study 2

MASSEY UNIVERSITY
TE KUNENGA KI PŪREHUROA
UNIVERSITY OF NEW ZEALAND

Confirmatory Factor Analysis and Validation of the SST and SST-B

INFORMATION SHEET

Description:

In this study, you will complete a survey that will ask a broad range of questions about your understanding of pieces of literary fiction and personal traits. Your participation should take no more than 55 minutes. Your participation is voluntary and you have the right to refuse to answer any question or withdraw from the study at any time without penalty. After you complete the survey you will be given an explanation of the study and provided with monetary compensation. You may also provide an email address which we will use to let you know about the findings of this research.

Participants should:

Be aged 16 or above. Not have a current neurological, psychiatric, or developmental disorder diagnosis. Not have severe visual impairment or epilepsy.

This study consists of an online questionnaire. The questionnaire consists of a number of multiple-choice and open-answer questions. Questions will be presented in a number of sections. If you choose to participate, please complete all sections in one sitting. You will not be able to resume at another time from where you left off. Your answers will not be saved until you complete all sections of the questionnaire.

Once the data have been analysed, we will ensure that we remove from the data set any information that might inadvertently include any identifying information. We will then make this non-identifiable data available to other researchers and might post it to an online repository.

If you wish to participate in this study and all of your questions have been answered, then please move to the next screen. If you do not wish to participate in this study, please return your submission on Prolific by selecting the 'Stop without completing' button.
Contact Information

Researcher
Joshua Robinson
Doctor of Clinical Psychology Student
Massey University
Palmerston North
New Zealand
Email: JoshuaResearch123@gmail.com

Supervisor
Dr Michael Philipp
School of Psychology
Massey University
Palmerston North
New Zealand
+64 6 951-8086
Email: m.philipp@massey.ac.nz

Supervisor
Dr Aaron Drummond
School of Psychology
Massey University
Palmerston North
New Zealand
+64 6 356-9099
Email: a.drummond@massey.ac.nz

Supervisor
Dr Clifford Van Ommen
School of Psychology
Massey University
Auckland
New Zealand
+64 6 951-0800
Email: c.vanommen@massey.ac.nz

Massey University School of Psychology – Te Kura Hinengaro Tangata
Palmerston North, New Zealand
T +64 6 3569-099 ext 85071 : W psychology.massey.ac.nz

This project has been evaluated by peer review and judged to be low risk. Consequently it has not been reviewed by one of the University's Human Ethics Committees. The researcher named in this document is responsible for the ethical conduct of this research.
If you have any concerns about the conduct of this research that you want to raise with someone other than the researcher(s), please contact Professor Craig Johnson, Director (Research Ethics), email humanethics@massey.ac.nz.

Next

APPENDIX D

Author Recognition Test Real Authors and Distractors

Real Authors				
Maya Angelou	Dick Francis	Toni Morrison	Isaac Asimov	Stephen King
Sidney Sheldon	Jean M. Auel	Judith Krantz	Danielle Steel	James Clavell
Robert Ludlum	J. R. R. Tolkien	Jackie Collins	James Michener	Alice Walker
Isabel Allende	F. Scott Fitzgerald	Vladimir Nabokov	Margaret Atwood	Sue Grafton
Joyce Carol Oates	Ann Beattie	John Grisham	Michael Ondaatje	Samuel Beckett
Ernest Hemingway	George Orwell	Saul Bellow	Brian Herbert	James Patterson
T. C. Boyle	Tony Hillennan	Thomas Pynchon	Ray Bradbury	John Irving
Ayn Rand	Willa Cather	Kazuo Ishiguro	Salmon Rushdie	Raymond Chandler
James Joyce	J. D. Salinger	Tom Clancy	Jonathan Kellerman	Jane Smiley
Clive Cussler	Wally Lamb	Paul Theroux	Nelson Demille	Harper Lee
Kurt Vonnegut	Umberto Eco	Jack London	E. B. White	T. S. Elliot
Bernard Malamud	Thomas Wolfe	Ralph Ellison	Gabriel Garcia Marquez	Virginia Woolf
Nora Ephron	Anne McCaffrey	Herman Wouk	William Faulkner	Margaret Mitchell

Retained Distractors				
Lauren Adamson	Mimi Hall	Eric Amsel	Robert Inness	Carter Anvari
Lilly Jack	Margarita Azmitia	Kirby Kavanagh	Reuben Baron	Frank Kiel
Christopher Barr	Stirling King	Gary Beauchamp	Susan Kormer	Reed Larson
Thomas Bever	Pricilla Levy	Lynn Liben	Elliot Blass	Dale Blyth
Alex Lumsden	Harrison Boldt	Hugh Lytton	Hilda Borko	Frank Manis
Jennifer Butterworth	Sophia Martin	Katherine Carpenter	Jennifer Marshal	Devon Chang
Morton Mendelson	James Morgan	Suzanne Clarkson	Ryan Morris	Charles Condie
Samuel Paige	Julia Connerty	Scott Paris	John Condry	Richard Passman

Edward Cornell	David Perry	Carl Corter	Peter Rigg	Diane Cuneo
Denise Daniels	K. Warner Schaie	Robert Siegler	Aimee Dorr	Frances Fincham
David Singer	Mark Strauss	Janice Taught	Tracy Tomes	Martin Ford
Nicole Waugh	Howard Gardner	Ava Wight	Sheryl Green	Noah Whittington
Frank Gresham	Ryan Gilbertson	Allister Younger	Steve Yussen	Carla Grinton

Omitted Distractors

Lena Johns	Oscar Barbarian	Lauren Benjamin	Brian Bigelow	Caleb Lim
Naomi Choy	Miriam Sexton	Geraldine Dawson	W. Patrick Dickson	Robert Emery

APPENDIX E

Case Study

Case Study 4: Research

Influences of my Doctoral Research on Clinical Practise

This case study was completed during the period of an internship as part of a Doctor of Clinical Psychology

In accordance with the Code of Ethics for Psychologists Working in Aotearoa/New Zealand the privacy of any clients is maintained by utilising pseudonyms and adapting identifying information.

Name: Joshua Robinson, Intern Psychologist, Ara Poutama Aotearoa, Palmerston North

Supervisor: Clifford Van Ommen, Senior Lecturer, Massey University, Auckland

Abstract

This case study focuses on the influences of my doctoral research on my practice as an intern psychologist. Initially, I provide a brief summary of relevant literature, my research objectives, and my methodology. This provides a framework to understand the subsequent reflections on how this research has helped to inform my clinical practice. Areas of focus will be the cognitive domains, with a particular emphasis on theory of mind; selection, scoring, and interpretation of psychometrics; using literature to inform practice; how to effectively use supervision; and how to problem solve and be adaptable.

Doctoral Thesis Overview

Rationale for Research

With recent estimates showing 67% of New Zealanders are regularly engaging with video games (Brand et al., 2017), understanding how this may be affecting cognition is key to allowing individuals to make informed choices regarding their media use. Much of the current research has focused on examining how violent video games may influence aggression (Drummond et al., 2018). Therefore, how video game play may be influencing other cognitive domains remains largely unclear. How video game play affects the core cognitive domains, as outlined in the Diagnostic and Statistical Manual of Mental Disorders, 5th Edition (DSM-5), was reviewed. Literature investigating the effects of video game play was particularly lacking regarding effects on Theory of Mind (ToM).

Given the paucity of research within this area, it is unclear whether existing media psychology theories provided an adequate framework for investigating relationships between video games and ToM. As such, further research is required to determine whether the General Aggression Model (GAM), the General Learning Model (GLM), or either model, is appropriate for investigating relationships between video game play and ToM.

Currently, measuring changes in ToM in neurotypical adults following an intervention (e.g., video game play) is impossible due to no alternate forms of any appropriate ToM measures existing. Therefore, there is a need to pilot an alternate form of a ToM measure to allow the measurement of changes in ToM in pretest-post-test experimental designs.

Aims

Aims in this current thesis were investigated across two studies. Additional exploratory analyses were also conducted which are not outlined below.

Study 1

The Short Story Task (SST), a newly developed ToM measure, theoretically measures a number of different facets of ToM (affective ToM, cognitive ToM, and reasoning ToM; Dodell-Feder et al., 2013). In addition, it does not display ceiling effects, a common problem observed when using ToM measures with neurotypical adults (Fitzpatrick et al., 2018). Thus, it is a viable measure from which to construct an alternate form, the Short Story Task B (SST-B). As such, the first aim was to:

1. To develop and pilot an alternate version of the Short Story Task with neurotypical adults.

The following psychometric properties were therefore examined:

- Concurrent Validity – The SST-B was correlated against the Reading the Mind in the Eyes Test – Revised Edition (RMET).
- Predictive Validity – The SST-B was correlated against the Autism Spectrum Quotient.
- Interrater Reliability – Scores between independent markers should be very strongly positively correlated.
- Alternate Form's Reliability – Intraclass Correlation Coefficients will be calculated between the SST and SST-B.
- Internal Consistency Estimations – These will be calculated utilising Cronbach's alpha.

The effects of engagement with single versus multi-player games on ToM has not been explored. As ToM is a facet of social cognition, it was hypothesised that individuals who engage with other video game players when playing would score higher on measures of ToM

than those who play in isolation. Further, the effects of engagement with different video game genres on ToM has not been explored. As such, the following hypotheses was generated:

2. To determine whether engagement with different video game genres is associated with differences in the ToM abilities of neurotypical adults.
    2.1. To determine whether the social context of video game play is related to differences in an individual's ToM abilities.

Further, Study 1 looked to examine whether existing media psychology theory was adequate for examining relationships between ToM and video games. As such, this thesis examined the final aims in Study 1:

3. To determine whether the GLM or GAM proposes a more accurate theoretical framework for investigating the influence of video game play on ToM abilities.
    3.1. To determine whether relationships between ToM ability and the social context of video games are consistent with predictions made by the GLM.

Study 2

Initially, Study 2 was going to follow a pretest-post-test experimental design whereby the SST and SST-B were to be used before or after a video game intervention. However, the COVID-19 pandemic prevented the conduction of any experimental research during 2020. Further, findings from Study 1 suggested that an experimental paradigm for Study 2 was unlikely to be fruitful. As such, the aims for Study 2 were adapted to the following:

Study 1 indicated that the SST and SST-B were not adequate alternate forms. However, the SST-B showed potential to be an improved measure of ToM as compared to the SST. Thus, the first aim of Study 2 was:

1. To determine whether the Short Story Task or the alternate form piloted during Study 1 is a better measure of ToM with neurotypical adults.

The following psychometric properties were therefore examined:

- Concurrent Validity – The SST and SST-B was correlated against the Reading the Mind in the Eyes Test – Revised Edition (RMET).
- Predictive Validity – The SST and SST-B was correlated against the Ritvo Autism and Asperger Diagnostic Scale 14 (RAADS-14).
- Interrater Reliability – Scores between independent markers should be very strongly positively correlated.
- Internal Consistency Estimations – These will be calculated using McDonald's omega
- While not fully preregistered, confirmatory factor analysis was used to examine the construct validity of these two measures.

During Study 1, the utility of the comprehension subscale in the SST and SST-B was also questioned. This is compounded by a growing desire for a reduction in the administration time of psychometrics, in both practise and research, as a cost-reduction method (Yates & Taub, 2003). Thereby, it was posited whether the comprehension subscales could be replaced with a simplified self-report measure of story understanding to reduce the SST and SST-B's administration time. Similarly, there is a growing desire to substitute tests of cognitive ability for short self-report questionnaires (Yates & Taub, 2003). Given that individuals can plausibly reflect and self-report upon their cognitive abilities (Craig et al., 2020), it was considered whether the ToM subscales of the SST and SST-B could be substituted with a simple self-report scale. Thereby, the second aim of Study 2 was:

1.1 To determine whether self-report scales of ToM ability and short story comprehension are adequate substitutes for tests of ToM ability and short story comprehension.

The Author Recognition Test (ART) is measure of familiarity with fictional print media (Stanovich & West, 1989). Higher scores on this measure are indicative of having read more literary fiction. Meta-analytic evidence suggests that reading literary fiction is associated with higher scores on traditional measures of ToM (Dodell-Feder & Tamir, 2018). Given the current limited number of studies examining this effect and the ongoing need to replicate findings within the psychological literature (Open Science Collaboration, 2015), the present study looked to replicate this. Thus, the following hypothesis was examined:

2. Replicate the finding that familiarity with literary fiction is associated with improvements in ToM abilities.

Methodology

Participants

Sample size was calculated using G*Power. A priori power analysis, set at .90 power to detect a medium effect size of r = .30 at the .05 alpha error probability, indicated that a sample size of 112 participants was required.

To be eligible to take part in Study 1, participants were required to:

1. Be actively playing video games for a minimum of three hours per week.
2. Be currently living in New Zealand, Australia, Canada, United States, or the United Kingdom.
3. Complete the survey on a desktop computer.
4. Be over the age of 16.

5. Not have a severe visual impairment.
6. Have no current or previous diagnoses of neurological, developmental, or psychological disorders.

Criteria 1 was to ensure video games were played with a high enough frequency to have a measurable effect on ToM if such an effect was present. Criteria 2 was implemented as ToM is a culturally bound construct (Oi et al., 2013). Therefore, ToM measures are likely to be invalid for individuals from non-western English-speaking countries where these measures were developed, piloted, and used. Criteria 3was implemented as some measures required use of a keyboard to complete. Criteria 4 was to comply with Massey University's code of ethical conduct. Criteria 5 was to ensure adequate perception of visual stimuli on the RMET. Criteria 6 was imposed as ToM is often impaired in the majority of these disorders (Cotter et al., 2018). Study 2 imposed criteria 2 through 6 with an additional criterion that participants had not participated in Study 1.

Procedure

All 112 participants, across both studies, were recruited online using Prolific. Prolific is an online survey hosting website where potential participants are informed about studies which they may then self-select to partake in. After reading through the information sheet and providing consent participants then completed the studies. The order in which participants completed the SST or SST-B across studies was counterbalanced to account for order and practise effects. Additionally, measures of traits associated with autism spectrum disorder (AQ and RAADS-14) were always completed last to account for any priming effects.

Internship Overview

The following section outlines reflections on my internship at Ara Poutama Aotearoa as of September 2021. Here, I worked primarily with clients in the Palmerston North community and Whanganui Prison. If required, I would also conduct assessment and treatment via Audio Visual Link in regions which did not have a psychology team (e.g., Taumarunui). Assessments primarily focused on determining an individual's risk of reoffending as well as any barriers to treatment engagement that would also require intervention (e.g., cognitive/neuropsychological functioning, mental health difficulties, personality functioning). Individual treatment was primarily with individuals who were at a high risk of reoffending, in line with the Risk-Need-Responsivity Model (Bonta & Andrews, 2016), and addressed dynamic risk factors for further offending (i.e., criminogenic needs). Additionally, treatment referrals might also require addressing aforementioned responsivity barriers to offence related treatment (e.g., low motivation due to low mood, personality traits and anxiety preventing engagement in a group treatment program).

Self-Reflections

The Cognitive Domains and Theory of Mind

My literature review initially focused on the cognitive domains within the DSM-5. This provided me with a strong foundation for neuropsychological assessment. When initially working with a client, this knowledge allowed me to screen for difficulties across a number of different areas. This is particularly important within a prison context given the prevalence of Traumatic Brain Injuries (TBI) and the diffuse difficulties often associated with this (Mitchell et al., 2017). When using neuropsychological tests, this information supported me to select tests that were appropriate to the difficulties the individual was presenting with.

Additionally, understanding of the integrated nature of these domains provided me with knowledge to screen for difficulties that might otherwise not be considered. For example, I had a client presenting with memory difficulties. Given my knowledge, I also screened for attentional and processing speed difficulties given these can impact on encoding. Given scores were lowered on subtests of attention and processing speed but not memory, this allowed me to make more nuanced recommendations regarding how treatment should progress with this individual than if I had only screened for memory difficulties.

Given my research focused on ToM, I also spent a significant amount of time researching this and other associated constructs (e.g., mentalizing, empathy). Understanding these constructs provided me with important foundational knowledge in a correctional setting. Specifically, it gave me an understanding of how attachment, mental health, trauma, and physical health (e.g., TBI) all link to these constructs and thereby how deficits in these may arise and manifest. Importantly, this helped me to understand and formulate an element of how some people may come to offend. For example, understanding that an inability to recognize and take someone's perspective, and the emotional detachment that can come with this, can reduce barriers to offending behaviours.

Psychometrics

A core part of my research was designing the SST-B. Part of this process involved the creation of a marking rubric, whereby a score of 0, 1, or 2 was assigned to a response based upon its accuracy. Close adherence to this rubric was crucial to ensure that accuracy between different raters was high (i.e., inter-rater reliability was high). A number of measures used at Ara Poutama, such as the Violence Risk Scale (VRS) and the Violence Risk Scale Sexual Offence Version (VRS: SO), operate on a similar basis; each item has a specific set of criteria to assign a score of 0 through to 3. Understanding the importance of close adherence to a

marking rubric, I set up a process whereby every time I scored these measures, I would read through the item description and how to assign each score. This meant that my total scores, when checked by my supervisors, have always aligned within ± 3 points of their scoring across the entire measure.

My research also highlighted to me the impact that context and culture can have on the psychometric properties of a measure. Unfortunately, few of the measures used at Ara Poutama have been validated with Māori and/or in a prison context. However, many of these tools still need to be utilized given their clinical utility (e.g., the Millon Clinical Multiaxial Inventory – Fourth Edition). As such, I always made sure to consider these limitations in their interpretation. Further, I would always research measures before administering them to determine whether they were invalid. On one occasion I considered administering the Test of Premorbid Functioning with a client of Māori descent. After researching the measure, I determined that this would not be appropriate (Dudley et al., 2017) and instead utilised qualitative indicators of premorbid functioning.

Literature Informed Practise

One of the most important skills that the research process imbued me with was the ability to find and evaluate literature and to then integrate this within my existing knowledge. I feel this has prepared me to work in any setting and has been particularly relevant to my work at Ara Poutama. It has allowed me to research areas and presenting problems I had little prior knowledge of (e.g., child sexual offending, Cluster A and C personality disorders) in a timely manner. Subsequently, I have then been able to integrate this new knowledge with my existing knowledge (e.g., Cognitive Behavioural Therapy principles) to create treatment plans and formulations. In essence, this skill has underpinned my ability to adapt my approach to meet and understand an individual's unique presenting problem throughout my internship.

Supervision

Supervision during my thesis was invaluable in preparing for supervision during my clinical work. Two years of doctoral thesis supervision allowed me to work out what learning strategies work best for me (e.g., being asked questions instead of being told what to do). This meant that I could convey this to my internship supervisors from the beginning thereby maximizing my learning. In addition, preparation for my thesis supervision sessions mimicked what they have looked like during my internship. Specifically, I would create a prioritized agenda across the week to ensure that the short time we had was utilized effectively and my biggest needs could be addressed and discussed. Finally, the thesis supervision process helped me to build resilience to criticism and to grow from this. Throughout my thesis supervision, my work was constantly critiqued and required adjustment. Understanding that this was part of the learning process has meant that I have been receptive to criticism from the start of my internship. Consequently, I have been able to utilize this to improve my practice instead of taking these critiques personally.

Adaptability and Problem Solving

Working on my thesis during the COVID-19 pandemic strengthened my ability to be adaptable and problem solve. Within three months I had to rework my entire second study to have new research aims/hypotheses that could be met through online surveying due to experimental research no longer being viable. Similarly, in clinical practice, I have been faced with many situations which have required me to utilize these skills. An intervention might not be working with a client; a client may not be engaging in their homework exercises; or the client might bring along a difficult experience that had occurred during the last week derailing my entire session plan. My ability to problem solve and be adaptable has

meant that, thus far, I have been able to overcome a diverse range of difficulties that this profession has presented me with.

References

Bonta, J., & Andrews, D. A. (2016). *The psychology of criminal conduct* (6th ed.). Routledge.

Brand, J. E., Todhunter, S., & Jervis, J. (2017). *Digital New Zealand 2018*. Eveleigh, NSW: IGEA.

Cotter, J., Granger, K., Backx, R., Hobbs, M., Looi, C. Y., & Barnett, J. H. (2018). Social cognitive dysfunction as a clinical marker: A systematic review of meta-analyses across 30 clinical conditions. *Neuroscience & Biobehavioral Reviews, 84*, 92-99. https://doi.org/10.1016/j.neubiorev.2017.11.014

Craig, K., Hale, D., Grainger, C., & Stewart, M. E. (2020). Evaluating metacognitive self-reports: Systematic reviews of the value of self-report in metacognitive research. *Metacognition and Learning*, *15*(2), 155–213. https://doi.org/10.1007/s11409-020-09222-y

Dodell-Feder, D., Lincoln, S. H., Coulson, J. P., & Hooker, C. I. (2013). Using fiction to assess mental state understanding: a new task for assessing theory of mind in adults. *PloS One, 8*(11), e81279. https://doi.org/10.1371/journal.pone.0081279

Dodell-Feder, D., & Tamir, D. I. (2018). Fiction reading has a small positive impact on social cognition: A meta-analysis. *Journal of Experimental Psychology: General, 147*(11), 1713-1727. https://doi.org/10.1037/xge0000395

Drummond, A., Sauer, J. D., & Garea, S. S. (2018). The infamous relationship between violent video game use and aggression: Uncharted moderators and small effects make it a far cry from certain. In C. J. Ferguson (Ed.), *Video game influences on aggression, cognition, and attention* (pp. 23-40). Springer.

Dudley, M., Scott, K., & Barker-Collo, S. (2017). Is the test of premorbid functioning a valid measure for Maori in New Zealand?. *New Zealand Journal of Psychology (Online)*, *46*(3), 72-79.

Fitzpatrick, P., Frazier, J. A., Cochran, D., Mitchell, T., Coleman, C., & Schmidt, R. (2018). Relationship between theory of mind, emotion recognition, and social synchrony in adolescents with and without Autism. *Frontiers in Psychology*, *9*, 1-13. https://doi.org/10.3389/fpsyg.2018.01337

Mitchell, T., Theadom, A., & Du Preez, E. (2017). Prevalence of traumatic brain injury in a male adult prison population and links with offence type. *Neuroepidemiology, 48*, 164-170. https://doi.org/10.1159/000479520

Oi, M., Tanaka, S., & Ohoka, H. (2013). The relationship between comprehension of figurative language by Japanese children with high functioning autism spectrum disorders and college freshmen's assessment of its conventionality of usage. *Autism Research and Treatment, 2013*, 1-7. http://dx.doi.org/10.1155/2013/480635

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science, 349,* Article aac4716. https://doi.org/10.1126/science.aac4716

Stanovich, K. E., & West, R. F. (1989). Exposure to print and orthographic processing. *Reading Research Quarterly*, 402-433. https://doi.org/10.2307/747605

Yates, B. T., & Taub, J. (2003). Assessing the costs, benefits, cost-effectiveness, and cost-benefit of psychological assessment: We should, we can, and here's how. *Psychological Assessment*, *15*(4), 478–495. http://dx.doi.org/10.1037/1040-3590.15.4.478