Spring 2021

Theory of Mind Measurements and Mechanisms: An Investigation of Construct Validity and Cognitive Processes in Theory of Mind Tasks

Ester Navarro Garcia
Claremont Graduate University

Theory of Mind Measurements and Mechanisms:

An Investigation of Construct Validity and Cognitive Processes in Theory of Mind Tasks

Ester Navarro Garcia

Department of Behavioral and Organizational Sciences

Claremont Graduate University

2021

A dissertation submitted to fulfil requirements for the degree of Doctor of Philosophy

Approval of the Dissertation Committee

This dissertation has been duly read, reviewed, and critiqued by the Committee listed below, which hereby approves the manuscript of Ester Navarro Garcia as fulfilling the scope and quality requirements for meriting the degree of Doctor of Philosophy in Psychology with a concentration in Cognitive Psychology.

Andrew Conway, Chair

Claremont Graduate University

Professor of Psychology

Kathy Pezdek

Claremont Graduate University

Professor of Psychology

Megan Zirnstein

Pomona College

Assistant Professor of Linguistics and Cognitive Science

Eleonora Rossi

University of Florida

Assistant Professor of Linguistics

Abstract

Theory of Mind Measurements and Mechanisms:

An Investigation of Construct Validity and Cognitive Processes in Theory of Mind Tasks

by

Ester Navarro Garcia

Claremont Graduate University: 2021

Understanding the perspectives of others is a critical skill. Theory of mind (ToM) is an essential ability for social competence and communication, and it is necessary for understanding behaviors that differ from our own (Premack and Woodruff, 1978). Although all individuals possess a ToM to varying degrees, bilinguals are especially adept to perspective-taking. Research has reported that bilinguals outperform monolinguals in ToM tasks (e.g., Goetz, 2003; Rubio-Fernandez & Glucksberg, 2012). However, the mechanisms underlying this effect are unclear. Studying individual differences in ToM performance between bilinguals and monolinguals can help explain these mechanisms. Yet this promising area of research faces an important challenge: the lack of psychometric research on ToM measurement. Recent research suggests that tests that measure the ToM construct might not be as reliable as previously thought (Warnell & Redcay, 2019). This hinders the interpretation of experimental and correlational findings and puts into question the validity of the ToM construct. This dissertation addresses these two questions empirically to improve our understanding of what constitutes ToM. Study 1 examines the structure of ToM, crystallized intelligence (Gc), and fluid intelligence (Gf) to understand (a) whether ToM constitutes a construct separate from other cognitive abilities and (b) to explore whether tasks of ToM present adequate construct validity. For this, three confirmatory factor analyses (CFAs) were conducted. The results demonstrated that a model with three latent factors

(ToM, Gf and Gc) did not adequately fit the data and was not significantly different from a model with only two latent factors (ToM-Gf and Gc). In addition, an exploratory factor analysis (EFA) showed that two of the ToM tasks loaded onto a Gf factor whereas one of the tasks loaded onto a third factor by itself. Finally, an exploratory network analysis (NMA) was conducted to observe relationships among the tasks. The results showed that the ToM tasks were no more related to each other than to some tasks of Gf and Gc, and that ToM tasks did not form a consistent cluster. Overall, the results of Study 1 suggest that ToM tasks are likely not measuring a monolithic ToM construct. Study 2 examines individual differences in metalinguistic awareness, executive function, and bilingualism as predictors of ToM. The results showed that all variables significantly predicted ToM, but bilingualism was not a significant moderator of ToM. Overall, the findings suggest that in this sample there was no difference in the processes used to predict ToM based on being bilingual or monolingual. Implications for measurement and individual differences in ToM are discussed.

Just for a moment, stand in their shoes.

– *President Joseph R. Biden, January 20, 2021*

Acknowledgements

I would like to thank my family and especially my mother, Maria del Mar García Úbeda, for her relentless support and encouragement, for nurturing my passion for learning, and nudging me towards my goals during my entire life. Este logro es tan tuyo como mío.

I would also like to thank my advisor Dr. Andrew Conway for introducing me to the topic of Theory of Mind and patiently teaching me the tools to complete this dissertation during the past five years, Dr. Kathy Pezdek for her immense help improving my research and writing skills and the members of my committee, Dr. Megan Zirnstein and colleague Dr. Eleonora Rossi for their continuous help throughout the process and kindness.

I would like to also thank my CGU colleagues and friends, especially Sara Goring, Dr. Erica Abed, and my research group, for being an endless source of wise advice, intellectual discussion, and friendship.

Finally, I would like to thank Dr. Giacomo Di Pasquale for accompanying me every step of the way and always be my other half.

This dissertation was funded by the 2020 Claremont Graduate University Dissertation Award.

Table of Contents

I. Introduction to Theory of Mind

1. The Conceptualization of Theory of Mind

How do humans understand how other people feel and what they believe? Psychologists and philosophers have long asked this basic question (see Wellman, 2017). Theory of mind (ToM) is the ability to understand the beliefs, knowledge, and intentions of others based on their behavior. The term was first coined by Premack and Woodruff (1978) to refer to chimpanzees' ability to infer human goals, and it was quickly adopted by psychologists to study humans' ability to infer and predict the behavior of others. This was followed by a vast number of studies on the topic. A simple search of the term "theory of mind" on PsycInfo reveals over 7000 articles and 1000 books on Theory of Mind. This is not surprising given that ToM is necessary for numerous complex cognitive tasks, including communication (e.g., Grice, 1989; Sperber & Wilson, 1995), criticism (Cutting & Dunn, 2002), deception (Sodian, 1991), joking and lying (Hughes & Leekam, 2004; Leekam & Prior, 1994), irony (Happé, 1994), pragmatic language competence (Eisenmajer & Prior, 1991), aggressive behavior (Happé & Frith, 1996), and problem solving (Greenberg, Bellana, & Bialystok, 2013). In addition, ToM has been observed across cultures and countries (Avis & Harris, 2016; Lee, Olson, & Torrance, 1999; Naito, Komatsu, & Fuke, 1994; Tardif & Wellman, 2000) and impaired ToM has been linked to psychiatric and developmental disorders, such as schizophrenia and autism spectrum disorder in both adults and children (Baron-Cohen, Leslie, & Frith, 1985, 1986; Frith, 2004; Hughes & Russell, 1993).

However, despite the numerous findings related to ToM (see Schaafsma et al., 2015), it is still unclear what the processes underlying ToM are. This might be partly due to the various operational definitions of the term ToM. For example, behavioral and neuroimaging research

usually distinguishes between language-independent implicit ToM (i.e., fast, automatic ToM) and culture and language-dependent explicit ToM (i.e., slower, deliberative ToM) (Apperly & Butterfill, 2009; Heyes & Frith, 2014; van Overwalle & Vandekerckhove, 2013). Other researchers instead distinguish between ToM as an emergent property based on experience and context and a latent ability that is expressed as the result of its interaction with general cognitive processes, such as working memory and executive function (Gopnik & Wellman, 1992, 1994, 2012; Leslie & Polizzi, 1998; Leslie, 1994). There are also differences between cognitive compared to affective ToM (Abu-Akel & Shamay-Tsoory, 2011; Poletti et al., 2012) and empathic ToM compared to representing the mental states of others (Preston & de Waal, 2002; Bernhardt, & Singer, 2012; van Veluw & Chance, 2014). Moreover, some researchers consider ToM in the realm of cognitive development while others refer to adult social cognition; more generally, some conceptions of ToM consider it the ability to understand the self as opposed to others, while other conceptions refer to ToM as empathic and emotional reactions. This wide variety of conceptual definitions suggests that the overarching concept of ToM as it is used by researchers in several fields likely entails a number of different processes and dimensions that represent different dimensions of a ToM ability (Schaafsma et al., 2005; Quesque & Rossetti, 2020). Thus, ToM research faces several challenges that need to be addressed to advance the field.

One consequence of the conceptual confusion around ToM research is that it hinders the creation of valid tests. That is, because the description of the processes underling ToM is confusing, it is difficult to find tests that adequately measure the processes that form ToM. The variety of terminology and the creation of a wide number of ToM measures with poor psychometric properties have contributed to the problem. Recent research shows that many of

the measures commonly used to assess ToM likely assess different processes (Warnell & Redcay, 2019), and it is unclear whether all of these processes really tap into an overarching ToM ability or whether they are tapping different lower-level processes (Quesque & Rossetti, 2020). In fact, there has been strong criticism of the way ToM is investigated and conceptually defined for a number of years (Bloom & German, 2000; Frith & Happé, 1994), yet the problem continues. Thus, one of the main goals of the field should be to address the lack of psychometric validity of ToM measures.

Another consequence of the conceptual confusion around ToM research is the lack of understanding of the processes involved in ToM. In particular, due to inconsistent terminology, instead of examining a general ToM ability, many studies have examined diverse subconstructs that might not completely represent the ToM ability. Schaafsma et al. (2015) suggested that one solution to the terminology issue is to not treat ToM as a "monolithic" ability (i.e., as an indivisible construct). Instead, researchers should consider the flexible nature of the construct when proposing theories that account for the processes that likely engage ToM. For example, Wellman (2018) proposed that one way to understand these processes is to examine populations that exhibit different ToM behaviors because of different individual experiences. Individual differences can help inform variation in achieving ToM milestones. For example, ToM seems to develop differently based on experience, such as different language ability (Milligan, Astington, & Dack, 2007), having knowledge of mental state words (Ruffman, Slade, & Crowe, 2002), having siblings, and growing up bilingual (Wellman, 2018). Regarding bilingualism, researchers have found that bilinguals on average complete ToM tasks at an earlier age than monolinguals (see Schroeder, 2019, for a meta-analysis). Thus, one goal of the field should be to address how various individual experiences including being bilingual impact ToM processes.

The current studies focus on two aspects of ToM. The first study examines the psychometric properties of current ToM measurements. The goal is to understand (a) the extent to which researchers are measuring adequately the ToM construct and (b) the measures that should be used, revised or abandoned by examining the tasks that load on a ToM construct, as opposed to other related but different constructs (i.e., verbal ability). This study is expected to shed light on whether ToM constitutes a coherent psychometric construct. The second study focuses on ToM research on bilinguals as a means of understanding whether ToM performance variation is the result of bilinguals engaging different processes than monolinguals. Specifically, the goal is to assess whether ToM performance can be predicted by different cognitive mechanisms for bilinguals than for monolinguals. Ultimately, the goal of the current research is to expand ToM theoretical frameworks by examining the extent to which the processes engaged in ToM vary based on individual differences.

1. Psychometric and Measurement Issues in ToM Research

Numerous psychometric tasks and tests have been created to measure ToM. The first task created to assess ToM was the false-belief task (Wimmer & Perner, 1983). This task could not have existed if it were not for the help of the philosophers who created the perspective-taking paradigm (Bennett, 2019; Dennett, 1978; Pylyshyn, 1978), inspiring Wimmer and Perner (1983). The wealth of ToM research that has followed Wimmer and Perner's study has led to the creation of a number of tasks and tests that assess different aspects of ToM. Some of the processes that these tasks measure include false belief understanding (Berstein, Thornton, & Sommerville, 2010; Wimmer & Perner, 1983), accounting for others' perspectives (Dumontheil, Apperly, & Blakemore, 2010), the ability to infer mental states from the expression of people's eyes (Baron-Cohen, Wheelwright, Hill, Raste, & Plumb, 2001; Baron-Cohen, Wheelwright,

Spong, Scahill, & Lawson, 2001), detection of faux pas (e.g., Baron-Cohen, O'Riordan, Stone, Jones, & Plaisted, 1999), deceptive intentions (e.g., Sebanz & Shiffrar, 2009), understanding others' thoughts (Keysar, 1994), and the difference between Level 1 perspective-taking (i.e., understanding that others' line of sight differs) and Level 2 perspective-taking (i.e., mentally adopting someone else's point of view) (Piaget & Inhelder, 1956; Samson, Apperly, Braithwaite, Andrews, & Bodley Scott, 2010), among countless others.

Despite the fast proliferation of ToM tasks, appropriate psychometric assessments of the validity of existing ToM measures have only recently been studied, with results suggesting concerns about the underlying structure being measured. Specifically, Warnell and Redcay (2019) examined coherence among ToM tasks using a psychometric approach. The researchers examined the relationship among different ToM measures (including the false belief task, the Reading the Mind in the Eyes test, and pragmatic language comprehension, among others) in child and adult populations. They found that an exploratory factor analysis did not support a clear structure underlying a ToM factor for the adult group. In addition, even though factor analysis was not possible due to low sample size for the child sample, correlations among the tasks administered to children also revealed poor correlations. These findings suggest that the measures used to assess ToM do not adequately tap into a reliable construct. The results of this study are consistent with recent theoretical accounts proposing that ToM is not likely a single construct, but that instead is a composite of both social and cognitive abilities (e.g., Apperly, 2012; Gerrans & Stone, 2008; Schaafsma et al., 2015). However, an earlier meta-analysis by Baker, Peterson, Pulos, and Kirkland (2014) found that the correlations among ToM tasks were generally higher than those found by Warnell and Redcay (2019). Therefore, further research is needed to understand whether measures typically used to assess ToM are indeed adequately

measuring the same underlying construct. Thus, it is critical to clarify what measures of ToM should be used, revised, or abandoned.

A first attempt to clarify ToM tasks was performed by Quesque and Rossetti (2020). The researchers conducted a systematic review of a large ToM task battery to assess the face validity of over 20 measures of ToM used by researchers from a variety of areas, including developmental, clinical, cognitive psychology, and cognitive neuroscience. They concluded that there were large differences in the underlying cognitive mechanisms that each of the tasks seemed to measure, including perspective-taking, eye tracking, and inference making. Importantly, they suggested that a paradigm shift in the methodologies traditionally used to explore social cognition are necessary to ensure terminological clarity and to advance the field. For this reason, they called for the need to identify and classify the measures that correctly assess ToM compared to others that likely only measure lower-order cognitive processes, such as kinematic processing (like automatic eye gaze movement; Obhi, 2012), social attention (Heyes, 2014) or emotion recognition (Oakley et al., 2016). Specifically, the researchers concluded that many of the tasks were likely measuring lower-order social-cognitive processes like those above, rather than a higher-order ToM ability, such as inhibiting one's perspective, creating models of alternative emotional responses, and updating one's own knowledge.

Further, Quesque and Rossetti (2020) emphasized the need for enforcing strict criteria for the use of ToM tasks. Specifically, they propose that any task that is used to assess ToM should meet two essential criteria: *mentalizing* and *nonemerging*. Mentalizing refers to whether success in a given task necessitates understanding others' mental states or whether it could be attributed to lower-order cognitive processes instead. For example, understanding emotion from people's eyes (Baron-Cohen, Wheelwright, Hill, Raste, & Plumb, 2001) might not actually tap into

higher-order processes required to understand how other people feel, but instead might be the result of lower-order perceptual responses. Nonemerging refers to whether a given task requires representing the mental state of another person when that mental state *differs* from the participant's mental state. Following the previous example, understanding the emotion expressed by somebody's eyes does not require that the participant inhibits their own emotion. That is, if the emotion in somebody's eyes represents anger, participants do not need to inhibit what they are feeling to realize it is anger. Higher-order ToM requires both understanding of others' mental states and the ability to inhibit one's own. Quesque and Rossetti (2020) argued that numerous tasks created and implemented to date do not meet both of these criteria and are therefore not measuring ToM.

Despite the apparent lack of construct validity across tasks, many studies have used ToM tasks to examine the relationship between ToM and other cognitive abilities. Specifically, research has shown that ToM is related to verbal ability and executive function (EF) (e.g., German & Hehman, 2006; Milligan, Astington, & Dack, 2007), leading many researchers to suggest that ToM performance requires the development of the processes underlying these abilities. However, given the poor correlations among tasks of ToM, it is unclear why the relationship between ToM and other cognitive abilities emerges. In fact, it is possible that the relationship among cognitive abilities, such as verbal ability and EF, is the result of ToM tasks that share processes with tasks of EF and verbal ability, rather than reflecting a relationship among constructs. Unsurprisingly, many of the ToM tasks used in the literature have components that, at face value, share processes with constructs commonly studied in the cognitive abilities research literature, such as crystalized and fluid intelligence. Therefore, to understand the construct validity of ToM measures, the relationship among ToM tasks and other cognitive

constructs should be examined. This would help elucidate the tasks that measure a specific ToM ability, and the tasks that measure other related, but different, cognitive abilities.

More specifically, the ToM research literature has largely overlooked the possible relationship between tasks of ToM and tasks of specific sub-abilities that constitute general intelligence. For over one hundred years, intelligence researchers have studied the ways in which people develop, use, and differ in cognitive abilities (for a review, see Kovacs & Conway, 2016). This long tradition emanates from research using cognitive test scores to extract a single common factor, *g*, representing general intelligence (Spearman, 1904, 1927). *g* is thought to be the result of the positive manifold, that is, the largely replicated finding that cognitive abilities are consistently positively correlated. For this reason, many researchers have traditionally interpreted *g* as the common cause underlying individual differences in task performance and the covariance among different measures (Gottfredson, 1997). Early intelligence research has shown that general intelligence is related to various specific abilities.

One of the earliest models of intelligence was the fluid/crystallized (Gf/Gc) model of intelligence (Cattell, 1963, 1971; Horn, 1994). The Gf/Gc model proposed that general intelligence was the result of two specific and opposite abilities: Gf or fluid intelligence and Gc or crystallized intelligence. Gf was defined as the ability to solve problems in novel situations, regardless of previous knowledge and Gc was defined as the ability to solve problems using previously acquired skills, largely related to the amount of formal schooling one has been exposed to (Kan, Kievit, Dolan, & van der Maas, 2011). These two abilities have been expanded and incorporated into more recent models of intelligence, including the Cattell–Horn–Carroll (CHC) model (McGrew, 2009), which combines the Gf/Gc model with other specific abilities, such as visual-spatial (Gv), processing speed (Gr) and memory retrieval (Gr). Importantly, Gf and Gc remain two of the

strongest factors in models of intelligence and have been replicated consistently across the literature and across neuroscientific and developmental studies.

While ToM has been related broadly to IQ, usually measured as school achievement (Baker et al., 2014; Coyle, Elpers, Gonzalez, Freeman, and Baggio, 2018; Dodell-Feder et al., 2013; Navarro, Goring, & Conway, 2021), it has not been psychometrically compared to specific measures of Gf and Gc. Gf and Gc represent two correlated but different dimensions of intelligence and several reliable tasks are used to measure each construct. It is important to understand whether the correlation between ToM and cognitive abilities is due to share variance among tasks that measure independent but related constructs, or whether, instead, existing tasks of ToM do not in fact measure a ToM construct but rather other cognitive abilities.

2. Current Theoretical Frameworks of ToM

Despite the methodological issues of ToM research, numerous theoretical approaches have been proposed to explain the processes underlying this ability. Specifically, ToM was first studied from a philosophical perspective, or "philosophy of mind" (Pylyshyn, 1978) that attempted to explain how people were able to "read" other people's minds (Davies & Stone, 1995). Unfortunately, philosophy of mind accounts were rarely empirically tested, forcing experimental psychologists to disengage from early theories (Apperly, 2010, p.5).

Empirically supported theories of ToM can be classified according to two features, namely (a) whether they describe ToM in terms of domain-general vs. domain-specific processes, and (b) whether they view ToM as a subset of cognitive "modules" (i.e., theoretical specialized compartments). Theories within (a) fall in the so-called Competence-Performance framework (Scholl & Leslie, 2001; Wellman et al., 2001); theories within (b) focus on describing

the presence of one or more related or unrelated modules (in which ToM's sub-abilities are divided) that form the ToM ability.

i. Competence framework

Gopnik and Wellman (1994)'s *theory-theory* is the main theory within the Competence framework. The theory-theory takes its name from the idea that children behave like "little scientists" (Gopnik, 1996a, p. 486) who create theories of people's intentions and revise those theories as new evidence emerges. The theory-theory proposed that ToM is an ability that emerges in childhood as a result of experience. Accordingly, children have a basic ToM (i.e., "folk psychology") to infer the mental states of others and they use it to naturally construct theories that explain the world around them. When children fail to reach a goal because they have not considered others' mental states, they adjust their theory accordingly. Therefore, children learn that individuals hold different mental states and that mental states can vary as a result of experience (Gopnik, Meltzoff, & Kuhl, 2000). This trial and error approach allows children to "realize" (i.e., through a conceptual change) that people have different mental states (Wellman & Liu, 2004). That is, children first have a "mentalistic" psychological theory based on non-representational states (e.g., desires and perceptions) and gradually develop a mental representational theory of other people's mental states (Flavell & Miller, 1998; Gopnik & Wellman, 1994; Perner, 1991).

However, the theory-theory assumed that this change occurs in childhood and therefore cannot explain findings showing that older children and adolescents sometimes make ToM errors if the difficulty of the task is age-appropriate (Miller, 2010), indicating that the older the individual the more complex the mental state can be. These led researchers to explore whether ToM was influenced by cognitive abilities that develop during childhood. Specifically, numerous

studies have found a relationship between EF and ToM, leading researchers to conclude that older children might require the use of more challenging tasks because EF increases with age. Many researchers propose that the strong relationship between EF and ToM (e.g., German & Hehman, 2006) and the fact that changes in ToM and EF seem to occur at about the same developmental stage (Carlson, 2005; Garon, Bryson, & Smith, 2008) indicate that ToM and EF are likely related (Carlson, Moses, & Breton, 2002; Carlson, Moses, & Claxton, 2004; German & Hehman, 2006; Perner, Lang, & Kloo, 2002; Hala, Hug & Henderson, 2003).

ii. Performance framework

To address the EF-ToM relationship, several theories have emerged within what is known as the Performance Framework. Specifically, Expression, Emergence, and Cognitive Complexity and Control-Revised (CCCR) performance theories attempt to describe how ToM develops by explaining how this development is affected by EF. All these theories have in common that they consider EF to be an essential aspect of ToM but differ in the specific role that EF plays in ToM use and development.

a. Expression

Expression theories suggest that an existing latent ToM is "activated" by EF and therefore can only be used when complex EF begins to develop. Leslie and collaborators (Leslie, 1994; Leslie, German, & Polizzi, 2005; Leslie & Polizzi, 1998) proposed the ToM mechanism (ToM-M), in which ToM was formed by a dual-component model, with a largely innate domain-specific ToM mechanism responsible for representing beliefs and desires, and a domain-general "selection processor" that develops gradually throughout the lifespan and allows interference resolution of conflicting perspectives via EF. Clearly influenced by nativist language production

theories (e.g., Chomsky's Language Acquisition Device, 1960), Leslie's theory suggests that a ToM system is in place from birth, but this system cannot be "expressed" until EF skills are available to control it (see also Carlson, & Moses, 1998). Thus, according to this theory, EF allows the "expression" of a latent ToM ability.

b. Emergence

Emergence theories suggest that EF allows the creation of an otherwise non-existing ToM, unlike Expression theory that proposes that ToM is an innate ability. Moses (2001) proposed that using EF (e.g., top-down self-control) makes understanding other people's mental states possible, even if they conflict with one's own mental states. Thus, EF processes can be abstracted to other contexts, evolving into an independent ToM mechanism that would otherwise not exist (Moses, 2001; Russell, 1996). A different twist of the Emergence theory proposes that, instead, developing ToM is what allows the development of EF (Perner, Lang, and Kloo, 2002). However, since temporal precedence cannot be established, this view of Emergence theory has not gained popularity.

c. CCCR

Finally, Zelazo, Muller, Frye, and Marcovitch (2003) proposed the Cognitive Complexity and Control-Revised (CCCR) theory. This theory suggests that both abilities, ToM and EF, are sub-abilities caused by an overarching general ability to reason about and attend to hierarchical rules. Specifically, Zelazo et al. (2003) propose that age-related changes in EF are the result of changes in the complexity of the rules that children can simultaneously use to solve a given task. According to this view, children solve coordinated conditional rules ("if I go to the store today, then I will buy milk, otherwise I'll drink juice") by reflecting on the rules these statements

represent, thus comparing them to other rules and embedding them under higher order rules. In this example, the conditional statement about the store is dependent on the completion of another event (today). As children age, this general rule-solving ability increases. Thus, CCCR theory proposes that both EF and ToM are byproducts of children's ability to follow and decide to follow hierarchical rules.

Most of the ToM theories can fit within the Competence-Performance frameworks. Nevertheless, more recent theories have focus on describing ToM from a different perspective. These theories can be considered parallel to the Competence-Performance frameworks; the main difference is that their focus is on describing the structure of ToM from a neurobiological perspective, while the Competence-Performance frameworks attempt to explain ToM from a developmental perspective. For this reason, more recent accounts are rooted in neuroimaging research and describe possible neural areas that contribute to ToM. At least four accounts about ToM have been proposed based on this evidence.

First, Gerrans and Stone (2008) proposed that ToM could be the result of sub-components that focus on different aspects of perspective-taking; more specifically, they proposed an overarching domain-specific ToM module (influenced by multiple low-level domain-specific social processes) and an overarching domain-general module (that interacts with domain-specific processes). This account attempts to explain ToM by describing the interrelated nature of ToM, EF, and contextual cues for resolution of domain-general and specific components of ToM. Second, Apperly (2012) proposed an account to unify ToM and the cognitive tasks used to measure it. This theory attempted to provide a better account of the psychometric structures emerging from ToM measurement. According to this account, ToM does not just constitute a specific construct as it was originally proposed, but instead spans multiple

cognitive abilities in an interactive way. Similarly, the account by Schaafsma et al. (2015) proposed that various independent domain-specific low-level processes (e.g., eye gaze, intention tracking) form a ToM construct, instead of having a single-module general ToM. In other words, Schaafsma et al.'s theory claims that ToM is formed by domain-general processes that explain relationships among tasks, but also have domain-specific components that are not accounted for by general processes. Some evidence from the last two accounts comes from neuroimaging studies showing that ToM is likely not just a single construct (Frith & Frith, 2003; Schurz, Radua, Aichhorn, Richlan, & Perner, 2014). Instead, despite general agreement over some of the areas engaged when responding to ToM tasks (i.e., ToM network), recent meta-analytic work has found that distinct activation profiles are found when examining separate tasks (as opposed to aggregated tasks) in a brain activation map (Van Overwalle & Baetens, 2009), suggesting that some areas are engaged more often when responding to some ToM tasks, but not others. In addition, brain activation patterns seem to also vary throughout the lifespan, with responses to ToM tasks starting off more diffused in early childhood and becoming more concentrated in adulthood (Bowman, Liu, Meltzoff, & Wellman, 2012; Bowman & Wellman, 2014). This evidence indicates that ToM is likely composed of different processes and can change throughout development.

Finally, Apperly and Butterfill (2009) proposed a theory to account for developmental differences throughout the lifespan. Specifically, Apperly and Butterfill's theory suggested that there is a two-system ToM ability that can account for both the EF-ToM relationship and conceptual changes based on experience or context. This dual-system view is based on classical dual-process theory that proposes that human cognition is defined by a distinction between effortless, intuitive, automatic processes (System 1) and effortful, deliberative, operational

processes (System 2) (De Neys, 2012; Evans & Stanovich, 2013; Kahneman, 2011; Pennycook, Fugelsang, & Koehler, 2015a). According to Apperly and Butterfill, in ToM System 1 is used by infants and young children, but also by adults when the situation does not require effortful processing, such as when there is no perspective conflict. System 1 thus precedes and contributes to System 2, a fully formed ability to comprehend other mental states that requires effortful processing.

Discerning among all of these theories is difficult because by definition, they are not independent, and there is evidence that supports several of the claims in each model. The reason why they are supported is likely because each theory focuses on different aspects of ToM. Some theories focus on the type of processes used (general vs. specific), some focus on the properties of psychometric tasks, some on brain regions engaged in ToM, and yet others focus on the developmental aspect of ToM. Therefore, it is possible to find evidence for each of these separate aspects, in turn supporting different theories. A unifying theory could bring all of this evidence together to explain ToM better from different angles. However, obtaining a unifying theory that encompasses all these areas is difficult. Much evidence comes from developmental research that has focused on examining ToM in child populations. Therefore, to better understand the origin of these theories, evidence from developmental studies should be considered first.

3. Developmental Evidence for the Competence and Performance Frameworks

Children's ability to understand mental states (e.g., beliefs, intentions, desires) is a foundational social-cognitive skill related to a variety of healthy developmental milestones, such as social competence, peer acceptance, and academic success (Carlson, Koenig, & Harms, 2013).

A vast amount of research has reported that by age 5 there are significant changes in children's understanding of mental states (Harris, 2006; Wellman & Liu, 2004). For example, by the end of their first year, children can treat individuals as agents with intentions (e.g., desires, goals) (Wellman, 2018). Specifically, Brandone and Wellman (2009) found that 6 and 8-month-olds have longer looking times to areas where they expect a person to look for an object than to areas where they do not expect a person to look and Behne, Carpenter, Call, and Tomasello (2005) found that 9-18 month-old infants were more impatient (e.g., reaching, looking away) when an adult could not hand them a toy than when an adult did not want to hand them the toy; this was not true for 6-month-olds. This behavior indicates that infants understand basic intentions by the time they are 9 months, but not earlier.

However, although children can execute many abilities that require basic perspective-taking by the age of 2 (i.e., emotion, intention, or perception), they largely do not understand mental concepts like knowledge and belief. Specifically, 1- and 2-year-old children often do not distinguish between their knowledge and beliefs and the knowledge and beliefs of others (Carlson, Koenig, & Harms, 2013). This was first demonstrated by Wimmer and Perner (1983). In their study, the researchers administered the Sally and Anne false-belief task[1] to children with ages ranging from 3 to 9 years of age. While most of the 5-9-year-olds provided accurate responses, the 3-4-year-olds did not, indicating that the ability to represent mental states of other

[1] The false-belief task is used to assess ToM in 2-5 year-olds (Wimmer & Perner, 1983). False belief understanding indicates that children comprehend (a) that agents have different intentions and knowledge, and (b) that thoughts can differ from objects in the real world (Wellman, 2018). The task presents two characters (e.g., Sally and Anne) in a child-friendly way. Children first see Sally hide an object in location A, then go away. While Sally is absent, Anne moves the object from location A to B. After children see the scene, they are asked whether Sally will first look for the object in location A or in location B. To respond correctly that Sally will look in location A, a child must infer that Sally does not know that the object has been moved, and therefore that she does not have the same knowledge and beliefs as the child. Children under 3 years of age generally fail to pass this task by answering that Sally will look for the object in location B, suggesting that they do not understand that mental states differ.

people becomes established at the ages of 4-6. Wellman Cross and Watson (2001)'s meta-analysis of 178 false-belief studies reported consistent findings: most 3-4-year-olds do not respond accurately to false-belief tasks compared to older children, indicating that they largely do not understand the mental states of others. Overall, research to date suggests that understanding mental states undergoes a change at age 3-4.

This change is largely distinguished by the difference between Level-1 and Level-2 perspective taking. That is, infants can understand that people see things differently (Level-1 perspective taking), even if they do not yet understand that others can think different things and have different perspectives (Level-2 perspective taking) (Flavell, 1974, 1977; Flavell, Everett, Croft, & Flavell, 1981). For example, Masangkay et al. (1974) administered a series of tasks to 2-to-5-year-olds (e.g., picture task, turtle task) in which objects presented a different perspective for the experimenter and for the children. They found that 2-year-olds correctly indicated when the experimenter could not see an object even when the child could (Level-1), but only older children indicated when the experimenter could see an object from a different perspective (i.e., from the top as opposed to from the left) than the child (Level-1). This suggests that Level-1 develops before Level-2. Similarly, Moll and Tomasello (2006) found that on average 24-month-olds, but not 18-month-olds, helped an adult find an object that was visible to them but not to the adult (Level-1), indicating that children younger than 24 months did not exhibit Level-1 perspective taking.

The developmental differences between Level-1 and Level-2 perspective have been largely taken to support theories within the Competence framework (e.g., theory-theory; Gopnik & Wellman, 1994). That is, children originally have a "theory" of what other people know, but since they are not always correct, they experience communication errors. This forces children to

adjust and reconstruct their initial theory to correctly understand what other people know, intend, and believe. Evidence from studies comparing Level-1 and Level-2 perspective taking is thought to indicate that ToM can evolve and become more sophisticated as a result of interaction with the world. However, other findings cannot be explained solely within the Competence framework. Specifically, some research has shown that resolution of false-belief tasks is related to executive functioning (EF) performance. This was first reported by Leslie and Polizzi (1998), who examined responses to false belief problems that required more EF, that is, negative false beliefs (i.e., a false belief task where the protagonist's desire is to *avoid* rather than approach a target). 4-year-olds in the study performed worse in the negative compared to the standard false-belief task, suggesting that more EF was needed for the negative tasks. This was extended by Carlson and Moses (2001), who conducted a correlational study to examine the relationship between EF (i.e., inhibitory control) and ToM in a sample of preschool-age children. The researchers found that inhibitory control was strongly correlated with ToM performance, even after controlling for factors like language, age, verbal ability, and family size. Numerous developmental and neuroscientific replications of these findings (Gerstadt, Hong, & Diamond, 1994; van der Meer, Groenewold, Nolen, Pijnenborg, & Aleman, 2011) have led to the conclusion that, unlike the Competence framework suggests, EF is a necessary factor for ToM development, and it likely allows the use of a complex ToM ability. Thus, research examining the relationship between EF and ToM provides support for theories within the Performance framework of ToM (e.g., ToMM theory; Leslie, 1994), that is, children can only utilize their latent ToM correctly when they develop EF naturally with age (i.e., when they are able to inhibit egocentric responses), but not before.

The conflict between these theoretical frameworks is known as the competence-performance debate (Wellman et al., 2001; Scholl & Leslie, 2001). Since both frameworks make similar predictions about ToM (i.e., that ToM begins developing in preschool years), it has not been possible to discriminate among them. To decide among these theories, researchers have studied how individual differences affect the development of ToM (Wellman, 2018). Specifically, growing up bilingual seems to help children reach the milestone of passing false-belief tasks earlier in development. Examining the reasons underlying bilinguals' performance can inform research on the extent to which these opposing frameworks explain ToM.

4. Explaining ToM mechanisms: bilingualism and ToM

The Competence-Performance debate (Wellman et al., 2001; Scholl & Leslie, 2001) cannot be easily resolved by studying standard samples of healthy children because both frameworks make the same predictions for this cohort. That is, theories from both frameworks propose that ToM develops between 3 and 5 years of age due to a different underlying process (i.e., experience and EF, respectively). However, Wellman (2018) suggested that studying individual differences in reaching the ToM milestone (i.e., passing false-belief tasks) could help researchers understand the processes underlying ToM performance. Individual differences, such as engaging in social-pretend play, having siblings, or growing up bilingual (Wellman, 2018) have been found to affect the development of ToM. Specifically, bilingual children (i.e., children who grow up learning and speaking more than one language) have been shown to outperform monolingual children in ToM false-belief tasks (Bialystok & Senman, 2004; Carlson & Moses, 2001; Goetz, 2003; Kovacs, 2009;) and this effect seems to be stable across tasks and not subject to publication bias (see Schroeder, 2019 for a meta-analysis).

Understanding the processes underlying bilinguals' ToM performance could help explain the processes that are engaged in ToM ability. Specifically, bilinguals present differences in factors that influence various cognitive mechanisms. For example, metalinguistic awareness and vocabulary size are predictors of ToM performance (Altman, Goldstein, & Armon-Lotem, 2018; Diaz & Farrar, 2017). Bilinguals also show different neurological development. Specifically, older adults who are bilinguals present stronger cognitive and linguistic efficiency (Baum & Titone, 2014). Differences in bilinguals' responses to cognitive ability tasks might reflect variation in the type of cognitive processes used by bilinguals as opposed to monolinguals. This indicates that individual differences in ToM also might be related to variation in the cognitive processes that bilinguals use when responding to ToM tasks.

Two explanations have been proposed for bilinguals' ToM performance. First, Bialystok and Senman (2004) proposed that bilinguals have enhanced EF as a result of constant conscious switching between languages (Bialystok, 1999). Specifically, Bialystok suggested that bilingual children have domain-general EF advantages over monolinguals on tasks that involve ambiguous and conflicting information thanks to their experience controlling both of their languages (Bialystok, 1999; Bialystok & Codd, 2000; Bialystok & Majumder, 1998; Bialystok & Viswanathan, 2009). Bialystok suggested that, because ToM also requires resolution of ambiguous and conflicting information, EF advantages might allow bilingual children to outperform their monolingual peers in tasks that require ToM. Thus, advantages in EF would increase ToM performance, supporting the *Performance* framework. Second, Goetz (2003) proposed, instead, that bilinguals' conscious switching between languages could be the result of bilinguals' awareness of the languages that people around them can and cannot speak (Kloo & Perner, 2003). This, in turns, could translate into improved metalinguistic awareness, that is,

awareness that objects and events can be represented in more than one way (Bialystok 1988, 1992, 1999), helping bilinguals comprehend that individuals have different mental states at an earlier age. Studies have found that young bilingual children are able to switch to the appropriate language of their interlocutor (Genesee, Boivin, & Nicoladis, 1996; Genesee, Nicoladis, & Paradis, 1995; Lanza, 1992), suggesting that bilinguals may be more aware of the fact that other people have different mental states, than are monolingual children (Goetz, 2003). Thus, advantages in metalinguistic awareness would increase ToM performance, supporting the *Competence* framework.

Both of these hypotheses (EF advantage and metalinguistic awareness) can potentially explain why bilinguals outperform monolinguals in ToM tasks and evidence for both of these views has been found. For example, Goetz (2003) examined ToM performance of bilingual and monolingual 3-4-year-olds in two temporally separate sessions. Goetz found that bilinguals performed better than monolinguals in most of the tasks in the first session, but the difference disappeared in the second session. Goetz proposed that this happened because bilinguals have more ToM "practice" as a result of their interactions with people who speak different languages, but this difference can be overcome if monolinguals practice their ToM, therefore suggesting that bilinguals have more ToM experience, but not enhanced EF ability.

On the other hand, Kovacs (2009) found results that supported the opposite view. Specifically, she administered 3-year-old monolingual and bilingual children two false-belief tasks. One false-belief task was a modified language-based task (i.e., requiring metalinguistic abilities) and the other was a standard false-belief task (i.e., not requiring additional metalinguistic abilities). Kovacs hypothesized that bilinguals should perform better in the modified language-based task than in the standard task because the modified task depicted a

language-switch context that should have facilitated ToM if bilinguals indeed have improved ToM skills due to metalinguistic awareness. However, Kovacs found that bilingual children outperformed monolinguals on both tasks, not just in the modified task, suggesting that bilinguals' performance is not due to a metalinguistic advantage, but instead could be due to a general EF advantage over monolinguals.

These contradictory results are reflective of the theoretical debate around the processes underlying the ToM ability. However, the idea that bilinguals might have an "advantage" in either EF or metalinguistic awareness has been rejected by some researchers. Instead, more recent studies indicate that bilingual children could be using different processes to engage ToM altogether, resulting in different development throughout the lifespan. In contrast to the Competence-Performance debate, these studies suggest that treating metalinguistic awareness and EF as dichotomous processes might not adequately account for bilinguals' performance. Specifically, Diaz and Farrar (2017) conducted a correlational study to examine whether bilinguals showed differences in the types of processes used to solve false-belief ToM tasks across a year. Matched children performed a false-belief task, a metalinguistic task, and an EF task. The researchers found that EF at time 1 largely predicted ToM performance at time 2 for monolinguals (but not bilinguals), while metalinguistic awareness at time 1 largely predicted ToM performance for bilinguals (but not monolinguals) at time 2. Similarly, Buac and Kaushanskaya (2019) found that EF predicted ToM performance for monolingual, but not bilingual, children while linguistic ability (measured using the Clinical Evaluation of Language Fundamentals, CELF; Wiig, Semel, & Secord, 2013 – often used to measure metalinguistic awareness in children) predicted ToM performance for bilingual but not monolingual children.

These findings suggest that instead of bilinguals having a quantitative advantage (i.e., bilinguals use the same cognitive processes as monolinguals, but they do so more effectively), bilinguals could have a qualitative advantage, that is, bilinguals and monolinguals might be using, to an extent, different mechanisms when completing ToM tasks with one set of mechanisms producing superior results. That is, bilinguals could rely on metalinguistic awareness to engage ToM more than other processes. By doing so, bilinguals might *alleviate* some of the cognitive load from (a) inferring the mental states of others and (b) inhibiting one's own mental states, which taxes EF resources, resulting in more accurate performance.

If the processes underlying ToM can vary based on specific individual experiences, such as being bilingual, then adults' performance should reflect these variations. Although early studies dismissed ToM research with adults because adults have a "fully developed" ToM (Apperly, 2010, p.86), there is evidence that adults show individual differences in ToM performance (Apperly, Back, Samson, & France, 2008; Apperly & Butterfill, 2009; Keysar, Lin, & Barr, 2003; Navarro, Macnamara, Glucksberg, & Conway, 2020). In addition, ToM develops gradually throughout the lifespan and becomes increasingly more accurate in adulthood (Dumontheil, Apperly, & Blakemore, 2010), suggesting that the processes used to engage ToM development in childhood could continue to be engaged in adulthood. In fact, compared to adults, older children and adolescents present neurological changes in brain areas engaged when responding to ToM tasks (e.g., right temporo-parietal junction), suggesting that ToM is not an immutable ability (e.g., Saxe, Whitfield-Gabrieli, Scholz, & Pelphrey, 2009). Because there is some evidence that adult bilinguals might also outperform adult monolinguals on ToM tasks (Javor, 2016, Rubio-Fernandez & Glucksberg, 2012, Navarro & Conway, 2021), it is possible that, just as it has been observed in children, adult bilinguals engage different processes than

monolinguals to perform ToM tasks. If this is the case, then studying bilingual and monolingual adults could shed some light on which specific processes are involved in ToM, and which processes bilinguals preferentially engage.

5. Individual Differences in ToM in Adulthood

In recent years, there has been an increase in the amount of research examining ToM in adults. This is likely because ToM is relevant for a number of everyday tasks performed by adults, such as complex social navigation, perspective taking, and complex communication (Sperber & Wilson, 1995, 2002). ToM understanding is first observed when children achieve the ToM *milestone* (Wellman, 2018) at 3-4 years of age, however this ability continues to develop along different dimensions throughout childhood (e.g., Carpendale & Chandler, 1996) and adulthood (Dumontheil, Apperly, & Blakemore, 2010). In fact, ToM seems to engage multiple brain areas throughout development, including the medial prefrontal cortex, and the left and right temporoparietal junction (i.e., the ToM network). In addition, different specific regions within these areas are utilized at different developmental stages, reflecting the way in which ToM processes change (Bowman & Wellman, 2014). For example, in infancy, regions engaged in ToM tend to be more diffuse (i.e., more areas are activated); however, there is a gradual incorporation of regions in the ToM network and a shift in the type of functions used as development proceeds (Bowman & Wellman, 2014). This suggests that changes that occur in infancy could influence later development, and therefore ToM development does not necessarily end in early childhood.

One example of developmental changes in ToM is reflected in research by Dumontheil, Apperly, and Blakemore (2010). The researchers examined participants aged 7 to 27 (divided in five age groups: 7-9, 10-11, 12-14, 15-17, and 19-27) on a ToM task that required taking into

account what a virtual avatar knew and did not know, compared to a control task where there was not an avatar. The researchers reported that ToM performance increased steadily with age, such that younger adults performed better than children and young adults performed better than all other groups. However, ToM errors were observed for all groups, suggesting that while ToM performance improves with age, since all groups presented ToM errors, ToM is still likely to be cognitively effortful and subject to individual differences. In addition, Dumontheil, Apperly, and Blakemore also found that while adults' ToM performance was better than all other groups, they did not perform better than the young adult group (aged 14-17) in a task of EF. According to the researchers, this might suggest that while ToM and EF are related, ToM continues to develop even after EF plateaus (for example, as a result of exposure to daily experiences where ToM is necessary). According to the researchers, the disassociation between ToM and EF suggests that ToM also relies on cognitive processes other than EF and that the type of process engaged at a given moment could vary based on an individual's cognitive "blueprint", such as being bilingual (Apperly & Butterfill, 2009; Keysar, Lin, & Barr, 2003).

While the specific processes that affect ToM performance in adults are unclear, research suggests that adults engage cognitively effortful processes to resolve ToM tasks. For example, a number of studies have reported that adults have egocentric biases about other people's thoughts and beliefs. Specifically, adults tend to think that other people will make decisions based on what *they* know but not necessarily what other people know. Mitchell, Robinson, Isaacs and Nye (1996) found that when participants knew that a character's belief was true, they judged it less likely that the character would change its mind than when the character's belief was false (i.e., *reality bias)*. Similarly, Birch and Bloom (2007) found that when participants knew the correct

location of a hidden object, they indicated that it was less likely that another person would look for the object in the incorrect location (i.e., *the curse of knowledge*).

These egocentric tendencies on perspective-taking tasks have been found in numerous studies among adult populations (Epley, Keysar, Van Boven, & Gilovich, 2004; Navarro, Macnamara, Glucksberg, & Conway, 2020; Nickerson, 1999; Royzman, Cassidy, & Baron, 2003), suggesting that engaging ToM is cognitively demanding, and therefore likely taxes EF resources. However, it is not clear whether adults can engage other processes, such as metalinguistic awareness, when utilizing ToM.

While an increasing number of studies have examined how task performance varies based on cognitive demands exerted by ToM, few studies have examined whether ToM performance in adult populations varies based on individual differences, just like it has been observed with children. The study by Diaz and Farrar (2017), described above, suggests that metalinguistic awareness is used to engage ToM by bilingual children, while EF seems to be more engaged by monolingual children. Early advances in metalinguistic awareness could influence normal ToM development, such that the processes engaged to use ToM early on in development could continue to be used throughout childhood and into adulthood, while other processes like verbal ability and EF might only be engaged when the task becomes more effortful. Researchers have considered that growing up bilingual merely helps children reach the ToM "milestone" earlier than monolinguals, but have not examined whether bilingualism has an impact on the processes engaged in ToM (Wellman, 2018). However, results of Diaz and Farrar (2017) and Buack and Kaushanskaya (2019) suggest that bilingual experiences can lead to using alternative processes, like metalinguistic awareness, to utilize ToM.

There is limited evidence that adult bilinguals outperform monolinguals in ToM performance. Rubio-Fernandez and Glucksberg (2012) found that college-age bilinguals had fewer eye fixations in the egocentric item of the false belief task than monolinguals, thus outperforming monolinguals. In addition, Rubio-Fernandez and Glucksberg found that the bilinguals in their study also outperformed monolinguals in the Simon task of inhibitory control. Finally, performance in the ToM task was correlated with performance in the Simon task for both groups. This led the researchers to conclude that one possible factor underlying bilinguals' ToM performance could be cognitive control. Javor (2016) also provided evidence of existing differences between bilingual and monolingual adults' ToM performance. Bilingual and monolingual adults completed a Hungarian version of the ToM short stories test (Dodell-Feder, Lincoln, Coulson, & Hooker, 2013) that requires participants to read several stories and indicate whether socially awkward or inappropriate situations occurred, as well as what the characters in the story felt, knew, and believed. Javor reported that, overall, bilingual participants outperformed monolinguals on accurate responses to the ToM test, suggesting that adult bilinguals might also outperform monolinguals in this test of ToM. Finally, Navarro and Conway (2021) found that bilingual adults outperformed monolinguals in responses to trials that required taking the perspective of another person and inhibiting their own perspective (i.e., director task). Overall, these findings suggest that bilingualism is associated with individuals' ability to take into account the perspective of another person, nevertheless it is unclear whether the processes involved in this advantage are the same processes found among children populations.

Given that bilingual children might use metalinguistic awareness (Buac & Kaushanskaya, 2019; Diaz & Farrar, 2017) to determine that others' perspectives differ from their own earlier than monolinguals, perhaps bilingual adults also engage different processes than their

monolingual peers to support ToM. This would indicate that ToM is more flexible than previously considered and that the use of a specific process (e.g., metalinguistic awareness) during childhood can carry over into adulthood. Examining this possibility would elucidate the extent to which ToM can be accounted for by the Competence and Performance frameworks as well as understanding the extent to which individual differences influence ToM.

6. Interim Summary

Theory of Mind (ToM) has been studied empirically for over 30 years, leading to a number of robust findings, including the age at which children begin showing belief understanding, the psychological disorders associated with impaired ToM, the relationship between ToM, language ability, and executive function, and the behaviors associated with ToM performance in adulthood. However, a number of methodological issues have recently arisen in the way ToM is conceptualized and measured (e.g., Quesque & Rossetti, 2020), suggesting that there could be deep issues in the construct validity of ToM tasks. In addition, there is still controversy about the processes that affect, intervene, and are engaged when using ToM, and little is known about how these processes can be affected or changed by individual differences, such as growing up bilingual.

This dissertation will focus on two areas that can contribute to better understanding ToM and ToM-related processes. First, the psychometric properties of the tasks will be examined. This is important (a) to ensure that ToM measures are assessing the two key criteria of ToM ability, that is, the mentalizing criterion and the nonemerging criterion (Quesque & Rossetti, 2020), and (b) to ensure that ToM tasks are not measuring other related but different constructs, such as fluid and crystallized intelligence. To do this, factor analysis and network modeling will be used

to assess whether ToM measures adequately represent a ToM construct, and to revise, maintain, or abandon measures that do not clearly assess ToM.

Second, individual differences in ToM performance will be examined to explore the processes underlying ToM. To this end, ToM performance and the processes that predict ToM performance will be studied among adult bilingual and monolingual populations. This will add to existing theories within both the Performance and Competence frameworks. Specifically, if bilinguals show that metalinguistic awareness (in addition to EF) can be used to predict ToM performance for the bilingual group, this would suggest that experience plays a role in the performance of ToM (supporting the Competence framework). Simultaneously, if monolinguals largely use EF, but not metalinguistic awareness, to engage ToM, then it would suggest that EF is also necessary for ToM performance. In other words, studying bilinguals could bridge both existing theoretical frameworks by showing the extent to which both frameworks can explain performance based on individual differences. Addressing both of these issues is a crucial step to further the field of ToM and to understand how humans decipher what other people think and believe.

II. Study 1: Psychometric Analysis of Theory of Mind Tasks

The goal of Study 1 is to examine the validity of ToM tasks by comparing performance on these tasks to measures of fluid intelligence (Gf) and crystallized intelligence (Gc) (Cattell, 1963). As mentioned, Gf and Gc are reliable constructs that predict a number of real-life outcomes and that represent related but different psychological attributes. Examining the differences between these constructs and ToM would allow us to explore whether the processes tapped by ToM tasks represent a unique ToM construct, or whether they instead measure other

related constructs. Recent research has reported that some measures of ToM reveal low inter-task correlations (Warnell & Redcay, 2019), suggesting that different tasks do not measure the same higher-order construct, but rather reflect task-specific processes (Quesque & Rossetti, 2020). For example, some ToM tasks require reading ability (e.g., Short Stories Questionnaire), while others require solving novel problems (e.g., Director Task). For this reason, it is necessary to understand whether these diverse tasks adequately assess the same underlying construct or whether they are actually measuring other abilities, such as Gf and Gc. For this purpose, participants completed a battery of ToM, Gc, and Gf tasks. If ToM tasks represent a distinct cognitive ability, then a three-factor model should best fit the data.

In addition, a psychometric network modeling analysis was conducted to examine the relationship among ToM, Gc, and Gf tasks. Psychometric network modeling conceptualizes cognitive abilities as interconnected networks composed of interactive processes (see Epskamp & Fried, 2018). This approach has many benefits. Specifically, psychometric networks are a powerful visualization tool to explore anticipated or unknown relationships amongst variables in a dataset and, unlike latent variable modeling, they are not constrained by the principle of local independence (i.e., the assumption that a latent factor causes any and all covariation among measures of the same construct). In addition, network modeling can account for the one-to-one relationships amongst tasks belonging to the same construct while at the same time estimating individual relationships between tasks belonging to different constructs. Finally, network modeling can estimate and plot associations between all observed variables, allowing investigators to describe and model current theories of ToM.

iii. Method

d. Design and Participants

An online sample of 208 participants was recruited using Amazon's Mechanical Turk (MTurk). The number of participants is based on the minimum sample size required for a three-factor Confirmatory Factor Analysis (CFA; Wolf, Harrington, Clark, & Miller, 2013). The inclusion criteria for the study were that all participants had to be based in the US and were over 18. Their ages ranged from 18 to 69 years old (M = 39.89; SD = 9.34, Median = 39). 116 participants identified as female. In terms of ethnicity, 148 participants identified as Caucasian, 13 identified as Black/African, 9 identified as Asian, 3 identified as Hispanic/Latino, and 7 identified as mixed ethnicity. None reported being color blind. In addition, all participants reported having correct-to-normal vision and were fluent in English. Only one person reported that English was not their native language. 13 participants reported speaking a language in addition to English fluently. The final sample size after outliers were removed was N = 203[2].

The design of the study was a correlational approach using two different psychometric modeling techniques. To conduct factor analyses, it is recommended that each latent construct includes at least three tasks. In this study, participants completed 9 tasks in total: 3 tasks of ToM, 3 tasks of Gf, and 3 tasks of Gc. Participants were randomly assigned to complete the tasks in one of three different orders. In order 1 (n =74), participants first completed the Gf tasks (Letter series, Number series, Ravens), followed by the Gc tasks (Synonyms, Antonyms, and General Knowledge) and by the ToM tasks (Director Task, RMET, SSQ). In order 2 (n = 58), participants

[2] Outliers are defined in the Data Cleaning section.

first completed the ToM tasks followed by the Gf and Gc tasks. In order 3 (n = 76), participants first completed the Gc tasks followed by the ToM and Gf tasks.

e. Measures

Theory of Mind Tasks

Reliability for all tasks was calculated using Cronbach's alpha. Three ToM tasks were used to assess the ToM construct. Even though traditional studies assume that ToM tasks tap into the same ToM construct, Warnell and Redcay (2019) found that the tasks vary substantially, even in terms of face validity. Therefore, it is important to investigate their construct validity. The ToM measures that were used in the study involve (a) taking the perspective of another person (i.e., Director task), (b) inferring mental states from people's eyes (i.e., Reading the Mind in the Eyes), and (c) interpreting socially inappropriate situations (Short Stories Questionnaire). See Appendices A-B for a sample of ToM tasks.

Director Task (Dumontheil, Apperly, and Blakemore, 2010; Legg, E. W., Olivier, L., Samuel, S., Lurz, R., & Clayton, N. S., 2017). The task was proposed by Keysar, Lin, and Barr (2003) and automated by Dumontheil, Apperly, & Blakemore, 2010. The current version was an automated version adapted from Legg et al. (2017) and run on Qualtrics. The task includes two conditions (Director, No Director) and 2 trial types (Experimental, Control). The stimuli are set up in a 4x4 shelf containing eight different objects arranged in different positions. In the Director Condition, an avatar called the Director is placed behind the shelf. Some of the compartments in the shelf are occluded from the Director's view so that only the participant can see those objects. The Director stands on the other side of the shelf and views the shelf from behind, so that only the objects in the open compartments are visible to the Director. The participant is then asked to attend to the instructions that the Director gives her in a speech box. On each trial, the Director

asks the participant to select one of the objects in the shelf (e.g., “the yellow sock”, “the small cup”). The participant responds by clicking on the correct object within the shelf. Participants have 5 seconds to respond to each instruction. Average accuracy and reaction times for all trials was recorded.

Conditions. In the *Director* condition, participants were asked to consider the perspective of the Director. To this end, participants were shown the shelf from the perspective of the director and were explicitly told that the Director cannot see objects in the occluded compartments. This condition assesses theory of mind because the participant has to remember that the perspective of the Director is not the same as theirs. In the *No Director* condition, participants are shown the same shelf, but the Director is not behind it anymore. Instead, participants are given a strategy; participants are told to ignore all objects placed in the slots with red backgrounds. This condition does not require theory of mind and instead requires the participant to inhibit prepotent information while keeping in mind a rule, therefore just requiring general executive function. The No Director condition is used as a control condition.

Trial types. *Experimental trials* are trials where the participant have to take into account the perspective of the Director. Participants have to select the correct response (i.e., the target), which is an object in the grid that both the participant and the director can see (the tennis ball in Appendix A), however in experimental trials the shelf also shows a competing object that can be the most appropriate response but only from the perspective of the participant (the golf ball in Appendix A-C). To respond correctly, participants have to consider the Director’s perspective and avoid clicking on the competing object that is only visible to them. In *Control trials*, the target object has a competitor but is always the best response from both perspectives and no competing object is included in one of the grey compartments (see Appendix A-C). *Filler trials*

referred to objects in the shelf that have no competitor and are visible to both Director and participant. The No Director condition included the same three type of trials. Different shelf displays or stimuli were created for the study. Each stimulus included three instructions, one of which was either an experimental or control, and two of them were filler trials. Experimental and control trials are never shown in the same stimulus. Control and experimental trials appear in a pseudorandom intermixed order throughout the task and the order of presentation of the stimuli is counterbalanced across participants. There are three written instructions per stimulus that were presented on a speech bubble near de Director (in the Director condition) or on the top right side of the shelf (in the No Director condition). Participants respond to a total of 16 control trials, 16 experimental trials and 64 filler trials in each condition. Participants also complete a practice trials before the Director condition.

Reading the Eyes in the Mind (RMET; Baron-Cohen, Wheelwright, Hill, Raste, & Plumb, 2001). The original paper based RMET task was programmed on Qualtrics. In the task, participants are presented with a series of 36 black and white photographs of the eye region of the face of White females and males of different ages (see Appendix B). One photograph is presented at a time and participants have no time limit to respond. Four words describing the potential emotion conveyed by the eyes are presented together with the photograph. Participants must select the word that best describes what the person in the photograph is feeling (e.g., sad, happy, scared, depressed). The test is thought to assess how well a person can understand other people's mental states. RMET scores range from 0 to 36 in a discrete fashion. Accuracy is measured in this task. The task lasts approximately 10 minutes.

Short Stories Questionnaire (SSQ; Dodell-Feder, Lincoln, Coulson, & Hooker, 2013; Lawson, Baron-Cohen, & Wheelwright, 2004). SSQ was implemented on Qualtrics. The test

contains 10 short stories, each divided into three sections. The stories involve utterances made by a character that could upset another character in the story (e.g., by incorrectly assuming someone's age). In this task, participants must infer the mental states of the characters (i.e., how they felt, what they thought). Because each story is divided into three sections, there are a total of 30 sections overall with at least four utterances in each section. 10 sections contained *blatant* target utterances (e.g., incorrectly estimating that a woman's age), 10 contained *subtle* target utterances (e.g., lying about remembering someone's name) and 10 contained *filler* control utterances (e.g., discussing the weather). Each section contained a corresponding question. The question asked the participant whether something said in the story could have upset someone. Participants had to judge whether the section contained an upsetting utterance and indicate what part of the text corresponded to the upsetting utterance. Each of the 10 stories included a filler question (i.e., a story that did not contain an upsetting question). The order of presentation of the stories is random. Participants are scored based on the number of targets identified. There are 10 stories, with three parts each and two of the three parts included either a blatant or subtle target utterance, resulting in 20 possible correct responses. Accuracy is measured in this task. Scores range from 0 to 20 in a discrete fashion. The task lasts approximately 15 minutes.

Fluid Intelligence Tasks

All fluid intelligence tasks were programmed in Qualtrics. The tasks are thought to measure the ability to follow rules and solve novel problems.

Letter Series (Ekstrom, French, Harman, & Dermen, 1976). In the task, ten sets of four letters are presented. All sets present 5 series of letters that followed a certain pattern except for one set. To respond correctly, participants must select the letter set that does not follow the pattern. Accuracy is measured in this task. The task automatically ends after 5 minutes.

Number Series (Thurstone, 1938). In the task, ten trials are presented showing a series of numbers of varying lengths in it. Each series of numbers is organized following a specific order or pattern. Participants are asked to select the number that would be consistent with the series from five choices. Accuracy is measured in this task. The task automatically ends after 5 minutes.

Raven's Progressive Matrices (Raven, 1938). A short version of Raven's figural inductive reasoning task was used to measure fluid intelligence. All items are divided into the even or odd items for a total 18 items per task from the Raven's Progressive Matrices Set II (Hamel & Schmittmann, 2006). Participants were randomly assigned to one of two task orders (odd trials or even trials). In this task, each item is part of a pattern of eight black and white figures arranged in a 3x3 matrix in which the last bottom right figure is missing. At the bottom of the matrix is a list of eight possible figures to choose from. Only one of those figures is the correct answer that best completes the pattern of the missing piece in the matrix. Figures range from simpler geometrical shapes to complex patterns. In each item there are a series of rules that the participant needs to find and keep in mind to find the right answer. Participants were given three practice trials before completing the task. A standardized score of correct responses is calculated. The task ends automatically after 15 minutes.

Crystallized Intelligence tasks

The crystallized intelligence tasks are programmed in Qualtrics. All tasks are thought to measure previously acquired knowledge.

Synonyms. The synonyms test presented participants with 10 words shown one at a time each with a list of possible answer choices. Participants had to choose the word whose meaning

was the same as the initial word displayed on the screen. Accuracy is measured in this task. Participants had 5 minutes to answer all 10 questions.

Antonyms. The antonyms test is identical to the Synonyms test, except that participants have to choose from the list of options the word that represents the opposite meaning to the word displayed. Accuracy is measured in this task. Participants have 5 minutes to answer all 10 questions.

General Knowledge. The general knowledge test consisted of 10 questions regarding general knowledge (e.g., "What planet is furthest from the sun?"). Participants have to type out their answers to respond and are asked to enter "I don't know" if they do not know the answer. Accuracy is measured in this task. Participants have 5 minutes to answer all questions.

f. Procedure

All tasks were administered via Qualtrics and participants accessed the study from Amazon's MTurk. Participants were assigned to one of three counterbalanced orders following an unbalanced Latin square design. Each order was counterbalanced based on the construct that the tasks measure (i.e., Gf, Gc, and ToM). That is, participants were randomly assigned to first complete the tasks of one of the three constructs, then completed the tasks of the second construct, and finally the tasks of the third construct., in a counterbalanced order. Tasks within each construct were presented always in the same order. Participants were allowed breaks in between tasks. Completing the battery of tasks takes approximately 90 minutes. Participants were compensated with $15.

iv. Results

Descriptive statistics for each measure and reliability estimates are presented in Table 1. Cronbach's alpha, a measure of scale reliability, was used to measure the internal consistency of the tasks in this study. All measures demonstrated adequate reliability according to Cronbach's alpha (i.e., $\alpha \geq .60$). In terms of the relationships among tasks, bivariate correlations between measures are reported in Table 2. As previous research has shown (Coyle et al., 2018; Warnell & Redcay, 2019), correlations among ToM tasks were low and all were more correlated with measures of Gf than with each other. In terms of Gc and Gf, the measures were overall moderately or strongly correlated. Letter series, number series and Ravens all presented correlations over $r = .30$ among themselves; general knowledge, synonyms, and antonyms were strongly correlated $r > .40$. Gc and Gf measures were also correlated as expected based on models of intelligence. Overall, Gc and Gf measures seemed to correlate adequately within their respective constructs. However, the ToM measures presented less clustered correlations. For example, the director task and SSQ presented low but significant correlations with all tasks, not just with ToM tasks, and the RMET seemed strongly related to the Gf measures in particular. To better understand these relationships, we conducted confirmatory factor analyses.

g. Data Cleaning

Missing Data and Outliers. The data were screened for missing values and outliers. The analysis indicated that only .05% of the data was missing. The missing data were mainly due to a technical issue in the Letter Series task that resulted in the loss of 16 responses. Values for the missing data were imputed using a multiple imputation-chained equation technique (via the *mice*

package in R; Azur et al., 2011) that uses Bayesian regression-based linear prediction to impute all of the missing data points.

Regarding outliers, univariate outliers were deemed negligible as the number of univariate outliers represented .04% of the data. Multivariate outliers were identified by generating Mahalanobis distance terms for each case (Tabachnick & Fidell, 2013). In total, 5 cases were identified as having a Mahalanobis distance greater than the associated critical value, (e.g., $\chi^2(31) = 61.09$) and were deleted list-wise.

Normality. Univariate and multivariate normality were assessed by examining skewness and kurtosis values and conducting several tests designed to assess multivariate normality. Prior to data imputation, no measures in the original dataset demonstrated violations to univariate normality due to extreme values of skewness (more extreme than ±3.00) and kurtosis (more extreme than ±10.00), as presented in Table 1. However, the multivariate normality assumption was violated, based on various tests (e.g., Mardia, Henze-Zirkler, Royston, and Zhou-Shao; see Alpu & Yuksek, 2016; Zhou & Shao, 2014). Following data imputation, multivariate normality was still not demonstrated across the multivariate normality tests used (all *p*s < .05), indicating that the data were largely non-multivariate normal. For this reason, factor analyses were conducted using an estimator adequate for non-normal data (i.e., robust maximum likelihood).

Homoscedasticity and Multicollinearity. Breusch-Pagan tests were conducted on the cleaned data, indicating that the residual variances were homoscedastic (*SSQ: BP = 8.3853, df = 6,p-value = 0.2112 RMET: BP = 8.684, df = 6, p-value = 0.1921; DT: BP = BP = 7.0446, df = 6, p-value = 0.3167*). Finally, for multicollinearity, the variance inflation factors (VIFs) for all of the variables were less than 5, indicating that the assumption has been maintained (James et al., 2014).

h. Confirmatory Factor Analysis

Confirmatory factor analysis is a technique used to test and estimate relationships among observed and unobserved variables to construct a measurement model. The measurement model can be used to assess whether tests that assess a construct are consistent with the theoretical definition of the construct of interest. To examine whether a construct is adequately being measured, the fit of the model to the data can be tested. The measurement model tests whether the observed variance-covariance matrix is equal to the variance-covariance matrix implied by the model. To decide whether a model fits the data, multiple fit indices are observed. Fit indices consider the fit of the model relative to the saturated model (where all relations are specified) or the null model (where no relations are specified). According to Kline (2015), adequate models should have a chi-square to degrees of freedom ratio lower than 2, a Comparative Fit Index (CFI) greater or equal to .90, a Standardized Root Mean Square Residual (SRMR) lower or equal to .08, and a Root mean square error of approximation (RMSEA) between .05 and .10 (Kline, 2015). In addition, factor loadings should also be observed; cognitive tests tend to present loading values of between .30 and .60. When several models are being compared, model comparison indicates whether the models are significantly different, indicating that one of the models represents the data more adequately.

In this study, CFA was used to assess the construct validity of the tasks by comparing model fit and loading paths. CFA requires the use of an estimation algorithm to compare iterated sets of values with the goal of minimizing the difference between the observed and the implied correlation matrix. Robust maximum likelihood is an adequate estimator for data that present multivariate nonnormality (Gibson & Ninness, 2005). Data from 203 participants were used. Three models were specified. The first model, Model 1, was a one-factor model where all

manifest variables were predicted by a single general construct. Model fit indices are in Table 3. Generally, Model 1 presented poor fit based on Kline's fit indices described above, with no indices within standard ranges. While the fit indices presented a poor model, the standardized factor loadings were overall adequate, with only Raven's Progressive Matrices presenting loadings under .30 (see Figure 2). This indicates that, as expected, a model with a single factor does not adequately represent the ability that the measures are thought to assess.

Model 2 was conducted next to examine whether the ToM tasks would be better represented by a Gf factor compared to a separate factor. Model 2 was a two-factor model where Gf and Gc were the latent factors. The tasks corresponding to the traditional ToM and Gf constructs were combined in this model based on their bivariate correlations. The reasoning behind this was to understand whether ToM tasks really do represent an independent construct or if they are rather more related to tasks of fluid reasoning. Model 2 originally presented a Heywood case, indicating a misestimation of the model. To avoid this issue and understand whether the ToM tasks adequately loaded into the latent factor, a model with only Gf and Gc (i.e., the classical two-factor model) was estimated and the estimates for each variable were used to constrain the Gf and Gc variables in Model 2. This avoided the emergence of a Heywood case and provided a more adequate representation of the ToM measures. Fit indices for Model 2 are in Table 3. Overall, Model 2 did not present an excellent fit based on Kline's fit indices, and no indices were within standard ranges, however some of the indices were close to excellent fit. Compared to Model 1, Model 2 presented better CFI, RMSEA, and chi-square to degrees of freedom ratio but slightly worse SRMR, while still outside the optimal range. Standardized factor loadings in Model 2 were adequate for the Gc and Gf factors, even though the ToM measures were loaded into the Gf factor. The correlation between Gf and Gc was strong, as it is

usually found (see Figure 3). These findings seem to indicate that although a two-factor solution was not a perfect fit for the data, nevertheless Model 2 seemed overall better than Model 1 and was not a complete misrepresentation of the data. Model comparison between Model 1 and Model 2 revealed that there was no significant difference between Model 1 and Model 2 (X^2 = -78.34, p = >.10).

Finally, Model 3 was conducted to examine whether a theoretically-driven three-factor model provided a more adequate representation of the data. Model 3 was a three-factor model where each set of tasks was grouped under the psychological construct they represent theoretically. Just like for Model 2, Gf and Gc tasks were constrained to the estimates reported in the model with only Gf and Gc to avoid a Heywood case. Fit indices for Model 3 are in Table 3. Contrary to what was expected, Model 3 did not present an adequate fit to the data based on fit indices. In fact, Model 3's fit indices were largely similar to those reported for Model 2, or slightly worse. Compare to Model 1, Model 3 overall presented overall a better fit. However, no indices were within adequate ranges. The factor loadings presented strong paths for Gf and Gc latent factors. However, the factor loadings for the ToM factor were poor (see Figure 4). None of them present loadings over .30 (see Figure 4). Model comparison between Model 2 and Model 3 revealed that there was no significant difference between the models and there was also no significant difference between Model 3 and Model 1 (X^2 = -75.19, p = >.10). One reason why the model presented poor fit might be due to the weak correlations among ToM tasks. Unlike the Gc and Gf tasks, the ToM tasks all presented poor loading paths and the correlations between the ToM factor and both the Gc and Gf factors showed correlations above 1, suggesting that perhaps the tasks in the ToM factor might have overlapping variance with some of the tasks in the other factors. This was further explored by examining modification indices (see Modified Model).

In general, none of the models presented excellent fit according to the fit indices. Overall, the two-factor model (in Figure 3) presented the most adequate indices, however there was no difference with the other models. Model 3 had slightly worse fit indices than the two-factor model but was not significantly different from Model 2.

i. Exploratory factor analysis

Given the poor fit of the models, we decided to conduct an exploratory factor analysis (EFA) to understand whether the data were indeed a good representation of the measurement model constructed in the CFAs. A parallel analysis was conducted to determine the number of factors that should be retained from the data. Parallel analysis is a method to determine the number of factors that the data form when conducting EFA. The analysis creates a random dataset with the same number of observations and variables as the original data and eigenvalues are computed for the randomly created dataset. Then, the randomly generated eigenvalues are compared to the observed eigenvalues. Because the random eigenvalues mostly represent random noise, only those factors that fall outside the random eigenvalues are considered real and are retained. The parallel analysis indicated that 2 factors should be retained (see Figure 5), rather than 3. This indicates that the third factor is likely so small that it is little more than random noise. However, to obtain a more interpretable EFA, we decided to follow the theoretical framework and extract three factors from the data, corresponding to the three psychological constructs. Because the data were not normal, the chosen estimator was weighted least squares and the rotation estimator was Oblimin[3], given the correlations among the variables. We

[3] Extraction techniques produce factors that are orthogonal and atheoretical. Rotation allows the transformation of the factor loadings, so they become more interpretable. Oblimin is an oblique (as opposed to

specified 3 factors in a factor analysis on all 9 measures. The results of the EFA are in Table 4. All variables with loadings greater than .30 were considered to load on a given factor. The results showed that all measures of Gc loaded adequately under the same factor. However, the Director task and RMET loaded under the Gf factor with the rest of the Gf measures, whereas the SSQ was almost entirely loading on the third factor by itself. These results suggest that in this sample, the ToM measures do not form a single construct, and the tasks seem to be measuring abilities closer to fluid reasoning, rather than a separate ToM construct. Given the results of the EFA, we decided to conduct an exploratory network model to better understand the relationships among measures.

j. Network Model Analysis

Exploratory Network Model Analysis (NMA) is an alternative analysis that conceptualizes cognitive abilities as interconnected networks composed of interactive processes (see Epskamp & Fried, 2018). In this technique, observed manifest variables are represented by *nodes* and estimated partial correlations amongst them are modeled via connections called *edges*. Therefore, this technique does not need the assumption of a superordinate unobservable factor. NMA can be used in conjunction or as an alternative to latent variable modeling and it presents a number of benefits. For example, because of its exploratory nature, NMA can be used as a powerful visualization tool to explore anticipated or unknown relationships amongst variables in a dataset. NMA is also not constrained by the principle of local independence. Unlike NMA, CFA is constrained by the principle of local independence. The principle of local independence

orthogonal) extraction technique, therefore it allows the factors to be correlated (which is often the case in psychological studies).

assumes that a latent factor causes any and all covariation among measures of the same construct, and therefore CFA does not allow to observe the variance that manifest variables potentially have in common. Instead, NMA estimates associations between observed variables without assuming that a latent cause is responsible for any and all covariation among measures of the same construct. For this reason, NMA can account for one-to-one relationships among nodes belonging to the same construct while at the same time estimating individual relationships between nodes belonging to different constructs. Finally, NMA estimates associations between all observed variables, therefore it is ideal for modeling cognitive theories that propose overlapping processes among processes within the same construct. In addition, since NMA is an exploratory technique, it can be used on the same data as the CFA.

In this study, NMA was used to examine whether tasks that assess ToM are adequately related to other ToM tasks and only slightly related to tasks that measure Gf and Gc. In addition, NMA was used to observe the relation between ToM tasks and Gf and Gc tasks to estimate the extent to which ToM relies on crystallized and fluid processes. For that purpose, tasks for all three constructs (i.e., ToM, Gf, and Gc) were included in the analysis and the parameters were set to the indices mentioned above. Data from the same 203 participants was used in this exploratory method. Based on the findings above and on previous research (Quesque & Rossetti, 2020), it was predicted that ToM tasks that do not meet mentalizing and nonemerging criteria such as SSQ, would present weak edges and would be more dispersed than tasks that meet these criteria. In addition, it was predicted that tasks that share construct validity would be more closely related, independently of the construct they assess theoretically.

NMA was conducted on the correlation matrix extracted from the dataset. The model was conducted and visualized using the *qgraph* package in R. The method and techniques used in this

study are consistent with recommendations from the network modeling tutorial written by Epskamp and Fried (2018). To conduct an NMA, two parameters must be set. Gamma is a hyperparameter that determines whether the model favors a more simple or complex structure per the number of estimated edges. Lambda is a tuning parameter that determines the rigorousness of removal of identified spurious edges that occur due to sample error. The NMA was generated using the graphical least absolute shrinkage and selector operator (gLASSO) regularization method to determine the level of network sparsity. Specifically, the extended BIC method was utilized, which produces simpler models, as gamma is automatically set to its most conservative setting (= .50). Consistent with Epskamp, Lunansky, Tio, and Borsboom (2018), lambda was be set to remove spurious (false-positive) edges while at the same time maintaining as many true edges as possible (i.e., .01). The settings used for the network model are designed to facilitate high-specificity during the estimation process, and high-sensitivity regarding network edge-pruning.

Figure 6 shows the results of the network model. First, both Gf and Gc measures show strong partial correlations and form two closely related but independent constructs. One of the tasks, Raven's, seems to have an especially central position in regard to the correlations among all three psychological constructs. While the Gf and Gc cluster together, the ToM tasks do not seem to represent a strong unified cluster. Even though the ToM tasks seemed to form a relatively solid construct in the CFA, in the NM they are visually less related to each other than the tasks that form the other constructs. In fact, they seem more related to other non-ToM tasks. Specifically, SSQ is slightly more related (.17) to Raven's Progressive Matrices than it is to either of the other ToM tasks (.1 and .06, respectively). In addition, the Director task and the RMET do not share any significant edges with each other, despite loading adequately on the

ToM latent factor in the CFA and on the EFA, suggesting that the relationship between the two measures that were observed in the EFA might be due to their relationship with Raven's. In addition, RMET seems to be more closely related to all the Gf tasks than to any other task, clustering with the Gf construct, rather than with the ToM construct, following the findings of the EFA. Overall, the ToM tasks do not seem to form a uniform construct separate from Gf and Gc, and rather seem to share processes with the tasks belonging to the other constructs than with each other.

In general, the results of the NMA show that these three ToM tasks are not as strongly related to each other as previously thought, thus questioning the overall construct that these measures assess. In addition, these findings replicate recent findings suggesting that there is little coherence among ToM tasks (Warnell & Redcay, 2019). Although the ToM tasks used in this study might be tapping on to *some* dimension of a ToM construct, these findings suggest that there are clear differences in the processes the tasks are assessing and that they are possibly measuring other cognitive abilities (such as Gf), rather than or in addition to just ToM. This indicates that more psychometric research is necessary to understand what tasks should be used to assess ToM in adults, but also to understand whether ToM should be interpreted as an independent monolithic construct, rather than a number of sub-constructs.

k. Modified Model

In addition to the above results, an additional CFA was conducted to examine the reasons behind the lack of fit in Model 3. For this purpose, modification indices were observed. Modification indices are estimates of the amount by which the chi-square value of a given model would be reduced, and therefore fit increased, if a specific parameter were modified in the model. That is, modification indices allow researchers to understand the ways in which the

model fit could improve based on a data-driven approach. Because of this, it is not advisable to use modification indices to specify a model a priori, but rather to examine potential issues in the existing model *a posteriori*.

To better understand the issues behind the misfit of Model 3, modification indices were observed. As it was also inferred from the EFA and NMA, the modification indices suggested that the fit of the model would improve if the RMET would be predicted by both Gf and ToM. These two modifications would considerably improve the fit of the model (see Figure 7). These modifications also improved the manifest variables loadings of the ToM latent factor, suggesting that the RMET is contributing variance to both constructs and therefore its use as a measure of purely ToM is dubious.

v. Discussion

The goal of Study 1 was to explore the psychometric properties of ToM tasks compared to fluid intelligence (Gf) and crystallized intelligence (Gc) (Cattell, 1963). As previous research has indicated, the ToM measures were poorly correlated (Warnell & Redcay, 2019), but presented adequate reliability. The CFA showed that none of the measurement models presented excellent fit. Specifically, Model 2 (the two-factor model with a Gf-ToM latent factor and a Gc latent factor) presented similar or better fit indices and path loadings than Model 3 (the model with Gf, Gc, and ToM), however they were not significantly different. These findings suggest that the tasks used to measure ToM might be more related to Gf tasks than to each other. In fact, the modified CFA model showed that the RMET shares processes with Gf and that a model where RMET was predicted by the ToM and Gf factors improved model fit. This was further confirmed by the exploratory factor analysis (EFA) in which the RMET and the Director task loaded under the Gf factor, whereas the SSQ loaded separately, indicating that the ToM tasks

tested here do not represent a unified construct. Finally, the NMA presented a visual description of the tasks. Specifically, the network model showed that measures of the well-established Gc and Gf constructs presented strong edges among the corresponding tasks (with weaker edges among the Gc tasks, representing the constructs' relationship) and overall clustered together. However, the edges of the ToM measures were weak, and the nodes were spatially closer to the Gf tasks (especially Raven's), than to each other. Specifically, the RMET seemed related to all Gf tasks but only presented a weak edge with the SSQ, and no edge with the director task. Similarly, the director task shared weak edges with measures of Gc and with the SSQ but not with the RMET, whereas SSQ presented weak edges with both ToM tasks and with Raven's. Overall, these findings suggest that in this sample of neurotypical adults, three of the most popular measures of ToM do not seem to reliably measure the same underlying construct.

III. Study 2: Examination of Processes that Predict ToM Performance

The goal of Study 2 was to compare ToM performance in bilingual and monolingual adults to test if the processes underlying ToM vary between groups. This study has potential to inform existing theories within both the Performance and Competence frameworks of ToM. Specifically, if bilinguals show that metalinguistic awareness (in addition to EF) is used to perform ToM tasks, this could suggest that experience-related processes play a role in the development of ToM (Competence framework). Similarly, if monolinguals largely rely on EF, with little to no influence of metalinguistic awareness, to complete ToM tasks, then it would suggest that, in addition to experience, EF is a key predictor of ToM performance. In other

words, studying bilinguals could bridge both existing theoretical frameworks by showing the extent to which both frameworks can explain ToM performance based on individual differences.

vi. Method

l. Design and Participants

An online sample of 186 participants was recruited using Amazon's Mechanical Turk (MTurk), 80 bilinguals and 106 monolinguals. The inclusion criteria for bilingual participants were that they were Spanish-English bilinguals, that they learned and used both of their languages before age 10, and that at the time of this study, they used both languages on a daily or weekly basis. The inclusion criteria for monolingual participants were that they only know English at a native level and have little knowledge of a second language. The total number of participants recruited was based on an *a priori* power analysis conducted in G*Power to determine the minimum sample size needed for a multiple regression analysis to have a 90% chance of detecting an increase in R^2 for a fixed model. The analysis indicated that a minimum of $N = 202$ participants (n=100) is needed for the study. All participants were based in the US. The monolingual group had a mean age of 37.52 (SD = 8.75, Median = 36.5) and the bilinguals had a mean age of 39.62 (SD = 12.75, Median = 37). 49 monolinguals and 38 bilinguals identified as female. None reported being color blind and all participants reported having correct-to-normal vision. All other demographic information is in Table 5a. The final sample size after outliers were removed[4] was N = 154, with 92 monolinguals and 62 bilinguals. The study is a correlational design where all participants completed a ToM task, an executive function task, and

[4] Outliers were determined based on the analyses conducted in the Data cleaning setion.

a metalinguistic awareness task composed of two subtests. In addition, all participants completed a survey including questions about their language use, culture(s) they identified with, and code-switching habits, among other demographic information (see Tables 5a-5f). Participants were compensated $10.

m. Measures

Theory of Mind

Due to the existing discussion regarding the validity of ToM measures (Quesque & Rossetti, 2020), in Study 2 the Director task was selected to assess ToM, as it is considered to assess both the mentalizing and nonemerging aspects of ToM (Quesque & Rossetti, 2020). In addition, the Director task was the only task that diverged from the Gf and Gc constructs in Study 1. The Director task is also thought to assess the perspective-taking component of ToM, rather than perceptual or emotional dimensions of ToM. This perspective-taking dimension has also been largely studied in adult non-clinical populations (e.g., Dumontheil, Küster, Apperly, & Blakemore, 2010; Ferguson & Cane, 2017; Pile, Haller, Hiu, & Lau, 2017; Samuel, Roehr-Brackin, Jelbert, & Clayton, 2019). For this reason, the same Director task described in Study 1 was used for Study 2. Task reliability was calculated by randomly splitting the observations in half and calculating Cronbach's alpha for each half of the dataset. Overall, reliability measured by Cronbach's alpha of internal consistency was 1.01.

Metalinguistic Awareness

Metalinguistic awareness was measured using the tasks developed by Cartwright et al. (2017) for adult samples. The tasks assess the contributions of metalinguistic awareness and cognitive flexibility. The two measures of metalinguistic awareness correspond to non-semantic

aspects of cognitive flexibility and are thought to assess the relative contributions of particular aspects of metalinguistic awareness and cognitive flexibility to differences between good and poor comprehenders. The overall reliability of the metalinguistic awareness measure was calculated using Cronbach's alpha for both of the subtests. Overall reliability was .73.

Graphophonemic awareness. The task consists of a 30-item Phoneme Counting Questionnaire in which participants have to count the phonemes in printed words (e.g., *filth* contains four phonemes). As mentioned above, Cronbach's alpha is a scale of internal consistency among tasks. Standardized item reliability based on Cronbach's alpha for this task was .87.

Syntactic awareness. This task consists of a 10-item word order correction task in which participants must reorder sets of words into syntactically appropriate sentences. Multiple solutions are possible for each set of words (e.g., "the words dog is small the timid" could be reordered as "The timid dog is small" and "The small dog is timid"). Scores are the total number of appropriate sentences generated across the ten sets of words. Standardized item reliability based on Cronbach's alpha reliability for this task was .87.

Executive function

Executive function allows the control of intentions and goals, while simultaneously avoiding interference. It is particularly relevant for a number of tasks, as it allows us to avoid automatic processes that create a conflict between a task and our own intentions. In Study 2, executive function was assessed with the Simon task.

Simon task (von Bastian & Souza, 2016). In this version of the Simon task, participants were presented with a circle on either the right-hand side or the left-hand side of the screen. In

each trial, participants were presented a fixation crossed for 250 ms followed by the circle. Participants were asked to press the left arrow key when the circle is green and the right arrow key when the circle is red. Congruent trials are trials where the green circle appears in the left side and the red circle appears on the right side (75% of trials) and incongruent trials are trials where the green circle appears in the right side and red circles appear on the left side (25% of trials). To ensure sufficient inhibitory control demands, only 25% of trials were incongruent (Logan & Zbrodoff, 1979). Participants responded to 200 trials. Accuracy and reaction time responses were recorded. Task reliability was calculated by randomly splitting the observations in half and calculating Cronbach's alpha for each half of the dataset. Overall alpha reliability was .97.

Verbal Ability

Semantic Verbal Fluency (**Binetti et al., 1996).** The same semantic verbal fluency task used in Study 1 was used to measure bilinguals' and monolinguals' verbal fluency. The task was modified for the bilingual group, so that it included four categories in Spanish in addition to the four categories in English (e.g., furniture, fruit, clothing, and animals). Presentation of categories was counterbalanced within each language and the order of language presentation was also be counterbalanced. Task reliability was calculated by randomly splitting the observations in half and calculating Cronbach's alpha for each half of the dataset. Overall alpha reliability was 1.03.

Bilingual Background

Participants completed a survey regarding participants' demographics, language history and use. The survey was composed of three well established questionnaires: Language Experience and Proficiency Questionnaire (Marian, Blumenfeld, & Kaushanskaya, 2007), the

Language and Social Background Questionnaire, and the Bilingual Switching Questionnaire (Rodriguez-Fornells et al., 2012).

n. Procedure

All tasks were administered via Qualtrics and participants had access to the study from Amazon's MTurk. Participants were randomly assigned to one of four counterbalanced orders to complete the three tasks and the questionnaire. The questionnaire was always the last item to complete, whereas the other tasks were counterbalanced for both monolinguals and bilinguals. The entire study took between 50-60 minutes to complete.

vii. Results

Descriptive statistics for each measure and reliability estimates are presented in Table 6. The metalinguistic test demonstrated adequate internal consistency according to Cronbach's alpha (i.e., $\alpha \geq .60$). In terms of the relationships among tasks, bivariate correlations between measures are reported in Table 7.

o. Data Cleaning

Missing Data and Outliers. The data were screened for missing values and outliers. This analysis indicated that there was no missing data. Regarding outliers, multivariate outliers were given priority as they are of greater concern than the less complex univariate outliers. Multivariate outliers were identified by generating Mahalanobis distance terms for each case (Tabachnick & Fidell, 2013). In total, 14 cases were identified as having a Mahalanobis distance greater than the associated critical value, (e.g., $\chi^2(31) = 61.09$) and were deleted list-wise. After

removal of the multivariate outliers, univariate outliers were deemed negligible as the number of univariate outliers represented .09% of the data. Thus, no other cases were removed.

Normality. Univariate and multivariate normality were assessed by examining skewness and kurtosis values and conducting several tests designed to assess multivariate normality (see Figure 8). Prior to data imputation, no measures in the original dataset demonstrated violations to univariate normality due to extreme values of skewness (more extreme than ±3) and kurtosis (more extreme than ±10), as presented in Table 6. However, multivariate normality was not demonstrated based multivariate normality tests (e.g., Mardia, Henze-Zirkler, Royston, and Zhou-Shao; see Alpu & Yuksek, 2016; Zhou & Shao, 2014). Following data imputation, multivariate normality was still not demonstrated across the multivariate normality tests used (all *p*s < .05), indicating that the data were largely non-multivariate normal.

Homoscedasticity and Multicollinearity. Breusch-Pagan tests were conducted on the cleaned data, indicating that the residual variances were not homoscedastic[5] ($BP = 30.255, df = 4, p - value = < .001$). Finally, for multicollinearity, the variance inflation factors (VIFs) for all of the variables were less than 5, indicating that the assumption was maintained (James et al., 2014). The violation of the homoscedasticity assumption was further examination by examining the histogram of the director task. The histogram revealed a bimodal distribution. For this reason, director task accuracy was divided using a median into a binary

[5] It is likely that the homoscedasticity assumption was violated due to a floor effect in responses to experimental trials from the bilingual group. As mentioned in the results section of Study 2, the data collected for the bilingual group were likely flawed due to reasons outside the design of the study.

variable. Therefore, a binary logistic regression was conducted to examine individual differences.

p. Group-level analyses

The main goal of the study was to examine individual differences in ToM between bilingual and monolingual adults. However, before conducting individual differences analyses, group-level analyses for each of the predictor variables were conducted to explore differences among the experimental conditions of each of the tasks as well as differences between bilinguals and monolinguals. Specifically, group-level analyses were conducted to ensure that the experimental manipulations in each task were successful. First, analyses were conducted to compare differences in accuracy based on trial type (experimental vs. control trials), condition (director vs. no director condition), and language group (bilingual vs. monolingual) in the director task with the goal of examining whether participants responded less accurately to critical trials in the director condition as well as whether there were differences by language group. Second, accuracy and reaction time were measured in the Simon task to explore whether there was a congruency effect (i.e., difference in responses based on congruent vs. incongruent trials) and whether responses varied by language group. Third, responses to both of the metalinguistic tests were compared for each language group to examine potential differences in accurate responses. Finally, responses to the Verbal Fluency tasks were analyzed to determine a) whether there were differences in verbal fluency in English for each language group and b) whether there were differences between English and Spanish verbal fluency for the bilingual participants.

Director task

Responses to the director task were examined to explore differences in responses between bilinguals and monolinguals for experimental compared to control trials for each condition. A 2

(trial type: experimental, control) x 2 (language group: bilingual, monolingual) x 2 (condition: director, no director) mixed factorial ANOVA was conducted on response accuracy. The three-way interaction was not significant, $F(1, 153) = .26$, $p = >.1$, $\eta_p^2 = .100$ (see Figure 9). However, there was a group by trial type interaction, $F(1, 153) = 74.21$, $p < .001$, $\eta_p^2 = .33$; bilingual participants responded less accurately to experimental than control trials (bilingual: $M = .23$, $SD = .35$; monolingual: $M = .74$, $SD = .37$). There was also a significant condition by trial type interaction $F(1, 153) = 39.46$ $p < .001$, $\eta_p^2 = .21$; participants responded less accurately to experimental compared to control trials in the Director but not in the No Director condition (bilingual: $M = .45$, $SD = .43$; monolingual: $M = .61$, $SD = .44$). These findings show that participants made fewer mistakes when they had to select items that both the participant and the director could see (i.e., control items) than when they had to select items that only the participant, but not the director, could see (i.e., experimental items). However, this only occurred when the participants completed the task in which they had to take the perspective of the director (i.e., Director condition) compared to when they merely had to follow a rule (i.e., No director condition). In addition, bilinguals seemed to largely underperform in experimental trials compared to control trials across both conditions, indicating that they largely responded inaccurately to items that the director could not see compared to monolinguals.

Simon task

Responses to the Simon task were examined to explore differences in responses between bilinguals and monolinguals for congruent compared to incongruent trials. To analyze responses to the Simon task (N = 154), the data were divided into two datasets based on reaction time and accuracy. Reaction time (RT) responses were filtered so that only correct responses were included. RT and accuracy each followed the same cleaning process described in the Method.

Both datasets were aggregated to obtain a mean RT and accuracy score per participant per condition (i.e., congruent, incongruent). Then, only participants with complete trials for both RT and accuracy were included. The total number of participants after this process was N = 144 for accuracy and N = 146 for RT.

First, a 2 (trial type: congruent, incongruent) x 2 (language group: bilingual, monolingual) mixed factorial ANOVA was conducted on RTs (see Figure 10). The predicted interaction was not significant, $F(1, 146) = .51$, $p > .10$, $\eta_p^2 = .003$. However, there was a significant main effect of group, $F(1, 146) = 17.41$, $p < .001$, $\eta_p^2 = .12$; bilingual participants had longer RTs than monolinguals (bilingual: $M = 733.41$ s, $SD = 269.97$; monolingual: $M = 575.40$ s, $SD = 154.85$) (see Figure 10). In addition, as predicted, there was a significant main effect of trial type, $F(1, 146) = 133.13$, $p < .001$, $\eta_p^2 = .48$; participants had shorter RTs when responding to congruent than incongruent trials (incongruent: $M = 658.76$, $SD = 221.74$; congruent: $M = 625.42$, $SD = 227.23$).

Next, a 2 (trial type: congruent, incongruent) x 2 (language group: bilingual, monolingual) mixed factorial ANOVA was conducted on accuracy (see Figure 10). Again, the predicted significant interaction was not significant, $F(1, 145) = .41$, $p > .10$, $\eta_p^2 = .002$. However, in the accuracy measure, there seemed to be a ceiling effect which might be responsible for the nonsignificant interaction. As predicted, there was a main effect of trial type, $F(1, 145) = 9.48$, $p < .001$, $\eta_p^2 = .39$; participants were more accurate when responding to congruent than incongruent trials (incongruent: $M = .94$ $SD = .05$; congruent: $M = .98$, $SD = .03$). The main effect of group was not significant.

Overall, the results showed that there was a congruency effect; all participants were more accurate and had shorter RTs when responding to congruent trials compared to incongruent. In addition, bilinguals again underperformed in this task, presenting longer RTs than monolinguals.

Metalinguistic task

Responses to the metalinguistic tasks were analyzed to examine differences in responses between bilinguals and monolinguals for each subtest. Analyses were conducted for each of the two metalinguistic subtests to examine any potential differences between bilinguals and monolinguals' performance. For this, two *t*-tests were conducted on the Graphophonemic and Syntactic Awareness tests. For responses to the Graphophonemic task, Levene's test showed that homogeneity of variance was not violated. An independent samples *t*-test indicated that there was a significant difference between responses of bilinguals and monolinguals, $t(215.93) = -5.82$, $p < .001$, $d = -.72$. Specifically, monolinguals more accurately identified number of phonemes than bilinguals (bilinguals: $M = 10.23$, $SD = 7.62$, monolinguals: $M = 15.32$, $SD = 6.81$) in the Graphophonemic test (Figure 11). Next, Levene's test showed that homogeneity of variance was not violated in the Syntactic Awareness test either. Another independent samples *t*-test was then conducted on the total number of correct responses to the Syntactic Awareness test[6]. There was a significant difference between responses of bilinguals and monolinguals, $t(237.43) = -8.81$, $p < .001$, $d = -1.05$. Specifically, monolinguals provided more grammatical sentences than bilinguals (bilinguals: $M = 6.43$, $SD = 4.22$, monolinguals: $M = 10.86$, $SD = 4.24$) in the Syntactic Awareness test (Figure 11). Together, these results suggest that bilinguals largely underperformed in both tests of metalinguistic awareness.

[6] As a reminder, the Syntactic Awareness test was open-ended, that is, participants could create as many grammatically correct sentences as possible. The Graphophonemic test had a range of 0-30 correct responses.

Verbal Fluency

Verbal fluency was reported to assess whether there were implicit differences in verbal ability between bilinguals and monolinguals, as well as differences in Spanish and English in the bilingual group. Responses to the verbal fluency task were averaged and examined in two subsequent analyses. The first analysis compares average verbal fluency in English for both bilinguals and monolinguals and the second test examines verbal fluency in Spanish compared to English for the bilingual participants. For this, two separate *t*-tests were conducted. In the first *t*-test comparing English verbal fluency of bilinguals to monolinguals, Levene's test showed that homogeneity of variance was violated, therefore a Mann-Whitney U test for nonnormal independent samples was conducted. There was a significant difference such that monolinguals outperformed bilinguals in the average number of words provided, $U = 4820$, $p < .001$, $d = -1.05$, indicating that monolinguals obtained a significantly higher score than bilinguals (monolinguals: $M = 12.83$, $SD = 4.02$, monolinguals: $M = 8.23$, $SD = 5.23$) in the English verbal fluency test (Figure 12). Next, another Levene's test showed that homogeneity of variance was also violated when comparing verbal fluency in Spanish among bilinguals. Another Mann-Whitney U test was conducted on the average verbal fluency score in English and Spanish. There was a significant difference between bilinguals' responses in English and Spanish, $U = 9540$, $p < .001$, $d = .87$. Specifically, bilinguals on average provided more correct words in English than in Spanish (English: $M = 8.23$, $SD = 5.23$, Spanish: $M = 4.49$, $SD = 3.32$) (Figure 12). These findings follow the same pattern of results reported for the other tasks above, suggesting that this group of bilinguals seems to have underperformed overall across all tasks. In addition, the finding that bilinguals had a substantially low score in Spanish and significantly lower than their

English score suggests that this group of bilinguals was in reality not composed entirely of bilinguals.

q. Individual differences analyses

Individual differences were examined next. The main goal of Study 2 was to explore whether metalinguistic awareness and EF are predictors of ToM, as well as whether bilingualism moderates their effect on ToM. To test this prediction, regression analyses were conducted with EF, Metalinguistic Awareness, and Language Group (bilinguals vs. monolinguals) as predictors, and ToM as the outcome variable (see Table 8 for descriptive statistics). As mentioned above, the director task (i.e., outcome variable) had a bimodal distribution. Therefore, the data violated the normality assumption and the homoscedasticity assumption for a traditional multiple regression. For this reason, responses to the director task were divided using a median split and a binary outcome variable was created where participants who scored below or equal to the median (Median = .40625) were assigned a 0 (i.e., incorrect) and participants who scored above the median were assigned a 1 (i.e., correct). Therefore, a binary logistic regression was used to analyze the effect of EF, metalinguistic awareness, and bilingualism on ToM. Bivariate correlations (Table 7) showed that ToM was significantly negatively correlated with RTs in the Simon task (r = -.37) and positively correlated with both metalinguistics tests (r = .31 and .38, respectively), thus indicating that it was adequate to perform regression analyses.

Three binary logistic regressions were conducted. To obtain a metalinguistic awareness score, the two metalinguistic tests were averaged and used as a metalinguistic awareness composite. Model 1 examined the effect of incongruent RTs in the Simon task and the metalinguistic composite as continuous variables, and Language group (bilinguals and monolinguals) as a categorical variable, on the two binary director task outcome. Model 2 was

identical to Model 1, but it also included the interaction between the Simon task and Language group and Model 3 was identical to Model 1 but including the interaction between the metalinguistic composite and Language group. Results for all models are reported in Table 8.

The estimates of Model 1 show that all predictor variables (the Simon task, language group, and metalinguistic awareness) significantly predicted performance on the director task. As a reminder, in a binary logistic regression, the estimate B represents the logit (i.e., the log odds). To make this estimate interpretable, the odds ratios and the probability of obtaining a 1 in the director task (i.e., the probability of obtaining a score above the median) for each predictor variable are also reported. The probability estimates suggest that participants with higher metalinguistic score and slower RTs have about 50% probability of obtaining a 1 (i.e., obtaining a score above the median) in the director task, and monolinguals have a 90% probability of obtaining a 1 in the director task. This is in line with the group-level findings that show that bilinguals underperformed across all tasks in this study. As Models 2 and 3 show, there was not a significant interaction between the Simon task or the Metalinguistic composite and language group, suggesting that in this study responses to the director task were not moderated by bilingualism. Figure 13 presents the results of Models 2 and 3.

The results of the individual differences analyses suggest that (a) bilinguals overall underperformed in the director task compared to monolinguals and (b) performance in the director task was overall predicted by performance in the Simon task and metalinguistic tests, as well as by being bilingual or monolingual. Specifically, slower RTs in the Simon task predicted more accurate performance in the director task, especially for monolinguals. In the metalinguistic awareness test, there was also a tendency for participants with higher metalinguistic scores to

also perform more accurately in the director task. These results are further discussed in the General Discussion.

viii. Discussion

The goal of Study 2 was to examine predictors of ToM performance in bilingual and monolingual adults. Specifically, whether metalinguistic awareness, EF, and bilingualism predicted ToM performance, as well as whether bilingualism moderated the effect of the other predictor variables on ToM. At the group-level, monolinguals largely outperformed bilinguals in all tasks. Importantly, bilinguals performed significantly worse in the verbal fluency task compared to monolinguals and bilinguals performed worse in the Spanish verbal fluency task than in the English verbal fluency task, suggesting that the participants in this sample were not bilinguals and might not have taken the study seriously. At the individual differences level, all EF, metalinguistic awareness, and bilingualism predicted ToM performance in this study. However, bilingualism did not significantly moderate the effect of either predictor variable. These findings are the first to suggest that that metalinguistic awareness and bilingualism are significant predictors of ToM. However, the disproportionate number of bilinguals who underperformed in this study suggests that there was a systematic problem with the bilingual sample recruited in this study. Given that data collection was conducted online, it is possible that the participants in the study were not in fact bilinguals or that they did not follow the instructions of the study, hindering the implications of the findings of Study 2.

IV. General Discussion

7. Study 1: The need for psychometric research of ToM.

The goal of Study 1 was to examine the psychometric properties of ToM tasks compared to fluid intelligence (Gf) and crystallized intelligence (Gc) (Cattell, 1963). As previous research has shown, ToM measures were poorly correlated (Warnell & Redcay, 2019), nevertheless, they presented adequate reliability. The confirmatory factor analyses (CFAs) showed that none of the models tested presented excellent fit. Specifically, the two-factor Model 2 (the model with a Gf-ToM latent factor and a Gc latent factor only) presented better fit indices and loadings than the three-factor Model 3 (the model with Gf, Gc, and ToM latent factors), even though the two models were not significantly different. These findings suggest that the ToM tasks tested in this study might be more related to Gf than to a separate construct. This was further confirmed by the exploratory factor analysis (EFA) in which the RMET and the Director task loaded under the GF factor whereas the SSQ loaded separately, suggesting measurement issues. In addition, the NMA presented a clearer picture of the relationships among the tasks. Specifically, the network model showed that, as opposed to the well-established Gc and Gf measures that presented strong edges and overall clustered together, the ToM measures were only poorly related and closer to the Gf tasks (especially Raven's), than to each other. Overall, these findings suggest that, at least in this non-clinical population of adults, the most popular measures used to test ToM across the literature do not seem to reliably measure the same underlying construct.

Recently, the director task has been the subject of intense debate. Specifically, researchers have questioned whether it measures a specific dimension of ToM or rather some other cognitive processes, such as mental rotation or selective attention (Rubio-Fernandez, 2017). In this study, the director task was poorly related to the other ToM constructs but

moderately correlated to the Gf and Gc tasks. It is unclear from this study whether the director task constitutes a better measure of a ToM construct than the other tasks measured in this study. It is possible that the director task represents a perspective-taking (cognitive), rather than social-cognitive or social-perceptual dimension of ToM, thus correlating more strongly with fluid reasoning (especially, Raven's progressive matrices that includes strong mental rotation and pattern seeking components). For example, research suggests that the SSQ is thought to be a social-cognitive measure whereas the RMET is thought to better capture social-perceptual ability (lower-order perceptual responses). Following this, it is possible that the director task represents yet another aspect of ToM. However, whereas the RMET and SSQ were moderately correlated, the director task presented poor correlations with the other two ToM tasks. This leaves unanswered the question of whether the director task represents a specific dimension of ToM that is only weakly related to other ToM dimensions, or whether instead the RMET and SSQ do not adequately capture ToM. Another possibility is that the director task does not necessarily measure a ToM-related ability but a different cognitive process. For example, Rubio-Fernandez (2017) has proposed that perhaps the director task could instead be measuring selective attention rather than solely ToM. Specifically, she proposes that the egocentric eye fixations often observed when participants see the critical items in the director task might not necessarily represent participants' egocentric fixations, but rather that these eye fixations might indicate that participants are instead using selective attention to discard the inappropriate item that hinders the listener's ability to carry out an action. This question is still unclear and requires further experimental research.

In general, these findings indicate that there are systematic issues around the conceptualization of ToM. Shaafsma et al. (2015) have pointed out that the use of ToM has been

"vague and inconsistent" across the literature and that there are deep inconsistences in the research, both related to the reliability of the measures and the claims made via the use of these measures in experimental designs. As discussed in the introduction, there are a number of different levels in which ToM is conceptualized and described, and different terminology is used to refer to the same construct; cognitive development, social cognition, self-understanding, perception of others, understanding logical inferences, emotion and/or empathy all seem to be included in the same umbrella term of ToM. However, it is unlikely that all these different aspects of cognition assess the same cognitive construct. As an example, in intelligence research, it is often thought that general intelligence encompasses a number of specialized sub-abilities (see Figure 14) whose positive correlations (i.e., positive manifold) represent a general intelligence construct. For each of these sub-abilities, a number of reliable tests (ideally at least three) are necessary to measure each sub-construct's validly and reliably. It is possible that ToM can form a similar construct with different but related sub-abilities (as suggested by Quesque & Rossetti, 2020). However, for this to be explored, researchers should take a step back and develop reliable and valid tests of *each* proposed sub-ability that they consider forms a ToM ability. In addition, this should be further examined and described for the different populations in which ToM is studied: clinical and healthy, adults and children. Only by conducting this important psychometric work would we be able to reconcile developmental, clinical, cognitive and neurological research on ToM.

8. Study 2: Individual Differences in ToM

The goal of Study 2 was to examine any differences in the processes that predict ToM performance in bilingual and monolingual adults. Specifically, the goal of Study 2 was to assess (a) whether being bilingual moderated the effect of metalinguistic awareness on ToM, and (b)

whether being monolingual moderated the effect of EF on ToM. Overall, the initial hypothesis was not supported. Monolinguals outperformed bilinguals in all the tasks. In addition, whereas EF, metalinguistic awareness, and bilingualism all predicted ToM performance in the individual differences analyses, bilingualism did not significantly moderate the effect of any predictor variable. It is worth noting that there was a disproportionate number of bilinguals who underperformed in the ToM task as well as the verbal fluency task. In addition, bilinguals performed worse in the Spanish verbal fluency task than in the English verbal fluency task. This indicates that there was likely a systematic problem with the bilingual sample recruited in this study. Due to the lack of control of the participants who completed the study online, it is possible that the bilingual participants were largely not bilinguals. In the raw data of the bilingual group, there was number of unreliable responses and fake responses (i.e., bot-generated responses), suggesting that non-human participants might have infiltrated the study. These issues overall compromise the implications of the findings of Study 2.

A positive outcome of this study is the high reliability observed among the metalinguistic tasks (Cartwright et al., 2017). There are not many metalinguistic tasks available for adult samples, therefore confirming the reliability of these tests allows for further research on the relationship of metalinguistic awareness and other cognitive abilities among adults. Metalinguistic awareness was also strongly correlated with executive function and performance in the director task in this study, suggesting that overall metalinguistic awareness is likely a relevant cognitive ability, even if it was not a moderating factor in this study. More studies examining individual differences in ToM should include a metalinguistic awareness task in addition to executive function measures.

Overall, Study 2 did not support the hypothesis that metalinguistic awareness is a stronger moderator than executive function in bilinguals compared to monolinguals. This provides support for the performance framework of ToM as opposed to the competence framework of ToM. Proponents of the performance view suggest that developing a more complex executive function alone is what allows children and, later, adults to successfully perform ToM tasks (Carlson & Moses, 2001; Bialystok & Senman, 2004; Leslie, 1994; Leslie, German, & Polizzi, 2005; Scholl & Leslie, 2001). Both metalinguistic awareness and EF are cognitive abilities that are developed in childhood and that improve throughout the lifespan, and both were significant predictors of ToM. This indicates that there are likely a number of cognitive abilities needed to complete ToM tasks and one's individual ability in each of these tasks will affect performance in ToM. However, these findings should be interpreted with caution due to the issues mentioned above.

Overall, Study 2 cannot shed light on the mechanisms that influence ToM performance. In addition to the issues related to the bilingual data, the implications of Study 1 suggest that it is necessary to understand what we are truly observing when we measure "ToM" before drawing inferences from experimental and correlational studies. The director task used in this study could represent perspective-taking, social-cognitive ability, selective attention, or just fluid reasoning. Given the low correlations among the ToM measures in general, tests and tasks should be better unified before examining individual differences in ToM if we want to have a better understanding of what influences this ability or abilities.

9. General Discussion

Overall, this dissertation poses a number of questions that warrant further study.

1. Is there one ToM or are there independent emotional, perceptual and/or cognitive processes underlying a ToM? Schaafsma et al. (2015) have posed the same question. The researchers raise the issue of what really constitutes an example of ToM. According to the researchers, the current conceptualization of ToM cannot be reduced to a number of basic processes and they wonder whether such a differentiated ability really exists: it "requires faith that there is indeed something distinctive about the core concept of ToM". As the researchers point out, neurological research has proposed that there are computational features (i.e.., domain-general) as well as content-specific features (i.e., domain-specific) that relate to people's ability to understand desires, intentions, and beliefs. Content-specific features need to be inherently social whereas computational features refer to the specific differentiated processes involved in this ability, such as decoupling, recursion, prediction, and causal inference (Schaafsma et al., 2015). Further, Quesque and Rossetti (2020) have proposed that there are likely several separate mechanisms that have been crammed under the term ToM. Specifically, they propose that the long list of tasks used to measure ToM correspond to different processes. For example, the RMET might assess "Facial Expression Categorization", the SSQ could be a task of "Mental States Ascription" and so on. To answer the question posed above, it is necessary that ToM measures reflect both general and specific computational processes used when one interprets desires, intentions, and beliefs in social contexts. Until this is achieved, it is not clear whether the field can move forward.

2. Are there individual differences in ToM performance? Can different subcomponents of ToM vary across individuals? Given the many similarities and relationship between ToM and other cognitive abilities, it is likely that ToM varies across individuals based on different factors, and it is also likely that the level of one's ToM ability can predict other life events. However, because

the formal conceptualization of the construct is so poor, it is not possible at this point to investigate such individual differences. While some research has looked into how ToM ability can vary across the lifespan (Dumontheil, Apperly, & Blakemore, 2010), it is not clear what these differences represent and whether they are accurate reflections of individual differences in ToM or, instead, of other related and/or overarching cognitive ability (e.g., EF).

In essence, this dissertation leaves more questions than answers, clearly reflecting the unstable state of the area of ToM research. These questions emphasize the need for a paradigm shift in ToM research where the theoretical research conducted to date is re-examined, the measurement and psychometric research revisited and improved, and the field is unified by taking into account the number of processes that are unique, as well as similar, to other well-established concepts.

References

Abu-Akel, A., & Shamay-Tsoory, S. (2011). Neuroanatomical and neurochemical bases of theory of mind. *Neuropsychologia*, 49(11), 2971-2984.

Alpu, O., & Yuksek, D. (2016). Comparison of some multivariate normality tests: A simulation study. *International Journal of Advanced and Applied Sciences*, 3(12), 73-85. doi: 10.21833/ijaas.2016.12.011

Altman, C., Goldstein, T., & Armon-Lotem, S. (2018). Vocabulary, metalinguistic awareness and language dominance among bilingual preschool children. *Frontiers in Psychology*, 9(OCT), 1–16. https://doi.org/10.3389/fpsyg.2018.01953

Apperly, I. (2010). Mindreaders. Psychology Press.

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

Apperly, I. A., & Butterfill, S. (2009). Do Humans Have Two Systems to Track Beliefs and Belief-Like States? *Psychological Review*, 116(4), 953–970.

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

Avis, J., & Harris, P. L. (2016). Belief-Desire Reasoning among Baka Children : Evidence for a Universal Conception of Mind. *Society for Research in Child Development*, 62(3), 460–467.

Azur, M. J., Stuart, E. A., Frangakis, C., & Leaf, P. J. (2011). Multiple imputation by chained equations: what is it and how does it work?. *International journal of methods in psychiatric research*, 20(1), 40-49.

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(1), 78–92. https://doi.org/10.1016/j.intell.2014.03.001

Baron-cohen, S., Leslie, A., & Frith, U. (1985). Does the autistic child have a "theory of mind"? *Cognition*, 21, 37–46.

Baron-Cohen, S., O'Riordan, M., Jones, R., Stone, V., & Plaisted, K. (1999). A new test of social sensitivity: Detection of faux pas in normal children and children with Asperger syndrome. *Journal of Autism and Developmental Disorders*, 29(5), 407-418.

Baron-Cohen, S., Wheelwright, S., Hill, J., Raste, Y., & Plumb, I. (2001). The "Reading the Mind in the Eyes" Test revised version: A study with normal adults, and adults with Asperger syndrome or high-functioning autism. *Journal of Child Psychology and Psychiatry and Allied Disciplines*, 42(2), 241–251. https://doi.org/10.1017/S0021963001006643

Baron-Cohen, S., Wheelwright, S., Spong, A., Scahill, V., & Lawson, J. (2001). Are intuitive physics and intuitive psychology independent ? A test with children with Asperger Syndrome. *Learning*, 5(January 2014), 47–78. https://doi.org/10.1111/j.1469-7610.2004.00232.x

Baron-Cohen, S., Leslie, A. M., & Frith, U. (1986). Mechanical, behavioral, and intentional understanding of picture stories in autistic children. Br*itish Journal of Developmental Psychology*, 4, 113–125).

Baum, S., & Titone, D. (2014). Moving toward a neuroplasticity view of bilingualism, executive control, and aging. *Applied Psycholinguistics*, 35(5), 857–894. https://doi.org/10.1017/S0142716414000174

Behne, T., Carpenter, M., & Tomasello, M. (2005). One-year-olds comprehend the communicative intentions behind gestures in a hiding game. *Developmental science*, 8(6), 492-499.

Ben-Zeev, S. (1977). The influence of bilingualism on cognitive strategy and cognitive development. *Child development*, 1009-1018.

Bennett, A. L. (2019). *An Empirical Longitudinal Analysis of Agile Methodologies and Firm Financial Performance.* ProQuest Dissertations and Theses. The George Washington University, Ann Arbor.

Bernhardt, B. C., & Singer, T. (2012). The neural basis of empathy. *Annual Review of Neuroscience*, 35, 1-23.

Bernstein, D.M., Thornton, W.L., & Sommerville, J.A. (2011) Theory of Mind Through the Ages: Older and Middle-Aged Adults Exhibit More Errors Than Do Younger Adults on a Continuous False Belief Task, *Experimental Aging Research*, 37, 5, 481-502, doi: 10.1080/0361073X.2011.619466

Bialystok, E. (1979). Explicit and implicit judgements of L2 grammaticality. *Language Learning*, 29(1), 81-103.

Bialystok, E. (1988). Levels of Bilingualism and Levels of Linguistic Awareness. *Developmental Psychology*, 24(4), 560–567. https://doi.org/10.1037/0012-1649.24.4.560

Bialystok, E. (1992). *Selective attention in cognitive processing: The bilingual edge*. Elsevier, 83, 501–513. https://doi.org/https://doi.org/10.1016/S0166-4115(08)61513-7

Bialystok, E. (1999). Cognitive complexity and attentional control in the bilingual mind. *Child Development*, 70(3), 636–644. https://doi.org/10.1111/1467-8624.00046

Bialystok, E., & Codd, J. (2000). Representing quantity beyond whole numbers: Some, none, and part. *Canadian Journal of Experimental Psychology/Revue Canadienne de Psychologie Expérimentale*. https://doi.org/10.1037/h0087334

Bialystok, E., & Majumder, S. (1998). The relationship between bilingualism and the development of cognitive processes in problem solving. *Applied Psycholinguistics,* 19(1), 69–85. https://doi.org/10.1017/s0142716400010584

Bialystok, E., & Senman, L. (2004). Executive processes in appearance-reality tasks: The role of inhibition of attention and symbolic representation. *Child Development*, 75(2), 562–579. https://doi.org/10.1111/j.1467-8624.2004.00693.x

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

Binetti, G., Magni, E., Padovani, A., Cappa, S. F., Bianchetti, A., & Trabucchi, M. (1996). Executive dysfunction in early Alzheimer's disease. *Journal of Neurology, Neurosurgery & Psychiatry*, 60(1), 91-93.

Birch, S. A. J., & Bloom, P. (2007). The curse of knowledge in reasoning about false beliefs: Research report. *Psychological Science*, 18(5), 382–386. https://doi.org/10.1111/j.1467-9280.2007.01909.x

Bowman, L. C., & Wellman, H. M. (2014). Neuroscience contributions to childhood theory-of-mind development. *Contemporary perspectives on research in theories of mind in early childhood education*, 195-224.

Bowman, L. C., Liu, D., Meltzoff, A. N., & Wellman, H. M. (2012). Neural correlates of belief-and desire-reasoning in 7-and 8-year-old children: an event-related potential study. *Developmental Science,* 15(5), 618-632.

Brandone, A. C., & Wellman, H. M. (2009). You can't always get what you want: Infants understand failed goal-directed actions. *Psychological Science*, 20(1), 85–91. doi:10.1111/j.1467-9280.2008.02246.x

Buac, M., & Kaushanskaya, M. (2019). Predictors of Theory of Mind performance in bilingual and monolingual children. International Journal of Bilingualism, 1367006919826866.

Carlson, S. M. (2005). Developmentally Sensitive Measures of Executive Function in Preschool Children. *Developmental Neuropsychology*, 28(2), 595–616. https://doi.org/DOI: 10.1207/s15326942dn2802_3

Carlson, S. M., & Moses, L. J. (2001). Individual Differences in Inhibitory Control and Children's Theory of Mind. *Child Development*, 72(4), 1032–1053. https://doi.org/10.1111/1467-8624.00333

Carlson, S. M., Koenig, M. A., & Harms, M. B. (2013). Theory of mind. Wiley Interdisciplinary Reviews: *Cognitive Science*, 4(4), 391–402. https://doi.org/10.1002/wcs.1232

Carlson, S. M., Moses, L. J., & Breton, C. (2002). How Specific is the Relation between Executive Function and Theory of Mind? Contributions of Inhibitory Control and Working Memory. *Infant and Child Development,* 11, 73–92. https://doi.org/10.1002/icd.298

Carlson, S. M., Moses, L. J., & Claxton, L. J. (2004). Individual differences in executive functioning and theory of mind: An investigation of inhibitory control and planning ability. *Journal of Experimental Child Psychology*, 87(4), 299–319. https://doi.org/10.1016/j.jecp.2004.01.002

Carlson, S. M., Moses, L. J., & Hix, H. R. (1998). The role of inhibitory processes in young children's difficulties with deception and false belief. *Child development*, 69(3), 672-691.

Carpendale, J. I., & Chandler, M. J. (1996). On the Distinction between False Belief Understanding and Subscribing to an Interpretive Theory of Mind. *Child Development,* 67(4), 1686–1706. https://doi.org/10.1111/j.1467-8624.1996.tb01821.x

Cattell, R. B. (1963). Theory of fluid and crystallized intelligence: A critical experiment. *Journal of Educational Psychology*, 54(1), 1-22.

Cattell, R. B. (1971). *Abilities: Their structure, growth, and action*. Houghton Mifflin.

Coyle, T. R., Elpers, K. E., Gonzalez, M. C., Freeman, J., & Baggio, J. A. (2018). General intelligence (g), ACT scores, and theory of mind:(ACT) g predicts limited variance among theory of mind tests. *Intelligence*, 71, 85-91.

Cutting, A. L., & Dunn, J. (2002). The cost of understanding other people: Social cognition predicts young children's sensitivity to criticism. *Journal of Child Psychology and Psychiatry and Allied Disciplines*, 43(7), 849–860. https://doi.org/10.1111/1469-7610.t01-1-00047

Davies, M., & Stone, T. (1995). *Folk psychology: The theory of mind debate*. In: M. Davies & T. Stone (Eds.). Blackwell.

De Neys, W. (2012). Bias and conflict: A case for logical intuitions. *Perspectives on Psychological Science*, 7(1), 28-38.

Dennett, D. (1978). *Beliefs about beliefs*. Cambridge.Org, 1(4), 568–570.

Diaz, V., & Farrar, M. J. (2018). The missing explanation of the false-belief advantage in bilingual children: a longitudinal study. *Developmental Science*, 21(4), 1–13. https://doi.org/10.1111/desc.12594

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).

Doherty, M., & Perner, J. (1998). Metalinguistic awareness and theory of mind: Just two words for the same thing? *Cognitive Development*, 13(3), 279-305.

Doherty, M.J. (2000). Children's understanding of homonymy: Metalinguistic awareness and FB. *Journal of Child Language*, 27, 367–392.

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

Eisenmajer, R., & Prior, M. (1991). Cognitive linguistic correlates of 'theory of mind' ability in autistic children. *British Journal of Developmental Psychology*, 9(2), 351–364. https://doi.org/10.1111/j.2044-835x.1991.tb00882.x

Ekstrom, R. B., Dermen, D., & Harman, H. H. (1976). *Manual for kit of factor-referenced cognitive tests* (Vol. 102). Princeton, NJ: Educational testing service.

Epley, N., Keysar, B., Van Boven, L., & Gilovich, T. (2004). Perspective taking as egocentric anchoring and adjustment. *Journal of Personality and Social Psychology*, 87(3), 327–339. https://doi.org/10.1037/0022-3514.87.3.327

Epskamp, S., & Fried, E. I. (2018). A tutorial on regularized partial correlation networks. *Psychological methods*, 23(4), 617.

Epskamp, S., Lunansky, G., Tio, P., & Borsboom, D. (2018). Recent developments on the performance of graphical LASSO networks [Blog post].

Evans, J. S. B., & Stanovich, K. E. (2013). Dual-process theories of higher cognition: Advancing the debate. *Perspectives on Psychological Science*, 8(3), 223-241.

Flavell, J. (1974). *The development of inferences about others*. In T. Mischel (Ed.), Understanding other persons. Rowman and Littlefield.

Flavell, J. (1977). *The development of knowledge about visual perception*. In Nebraska Symposium on Motivation, 43–76.

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

Flavell, J., & Miller, P. (1998). *Social cognition*. In W. Damon (Ed.), Handbook of child psychology (Vol. 2, pp. 851–898). John Wiley & Sons Inc.

Friesen, D. C., & Bialystok, E. (2012). Metalinguistic ability in bilingual children: The role of executive control. *Rivista di psicolinguistica applicata,* 12(3), 47.

Frith, C. D. (2004). Schizophrenia and theory of mind. *Psychological Medicine,* 34(3), 385–389. https://doi.org/10.1017/S0033291703001326

Frith, U., & Frith, C. D. (2003). Development and neurophysiology of mentalizing. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 358(1431), 459–473. https://doi.org/10.1098/rstb.2002.1218

Garon, N., Bryson, S. E., & Smith, I. M. (2008). Executive Function in Preschoolers: A Review Using an Integrative Framework. *Psychological Bulletin,* 134(1), 31–60. https://doi.org/10.1037/0033-2909.134.1.31

Genesee, F., Boivin, I., & Nicoladis, E. (1996). Talking with strangers: A study of bilingual children's communicative competence. *Applied Psycholinguistics*, 17(4), 427–442.

Genesee, F., Nicoladis, E., & Paradis, J. (1995). Language differentiation in early bilingual development. *Journal of Child Language,* 22(3), 611–631.

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

Gerrans, P., & Stone, V. E. (2008). Generous or parsimonious cognitive architecture? Cognitive neuroscience and theory of mind. *British Journal for the Philosophy of Science*, 59(2), 121–141. https://doi.org/10.1093/bjps/axm038

Gerstadt, C. L., Hong, Y. J., & Diamond, A. (1994). The relationship between cognition and action: Performance of children 3 1/2-7 years old on a Stroop-like day-night test. *Cognition*, 53(2), 129-153.

Gibson, S., & Ninness, B. (2005). Robust maximum-likelihood estimation of multivariable dynamic systems. Automatica, 41(10), 1667-1682.

Goetz, P. J. (2003). The effects of bilingualism on theory of mind development. *Bilingualism: Language and Cognition*, 6(1), 1–15. https://doi.org/10.1017/s1366728903001007

Gopnik, A. (1996). The scientist as child. *Philosophy of Science*, 63(4), 485-514.

Gopnik, A., & Wellman, H. (1994). *The 'theory' theory*. In L. C. John Tooby, Alan M. Leslie, Dan Sperber, Alfonso Caramazza, Argye E. Hillis, Elwyn C. Leek, Michele Miozzo (Ed.), Mapping the Mind: Domain Specificity in Cognition and Culture. Cambridge: Press, Cambridge University.

Gopnik, A., Meltzoff, A., & Kuhl, P. (2000). *The scientist in the crib: What early learning tells us about the mind.* William Morrow Paperbacks.

Gottfredson, L. S. (1997). Why g matters: The complexity of everyday life. *Intelligence*, 24(1), 79-132.

Greenberg, A., Bellana, B., & Bialystok, E. (2013). Perspective-taking ability in bilingual children: Extending advantages in executive control to spatial reasoning. *Cognitive Development*, 28(1), 41–50. https://doi.org/10.1016/j.cogdev.2012.10.002

Grice, H. (1989). *Studies in the Way of Words*. Harvard University Press.

Hala, S., Hug, S., & Henderson, A. (2003). Executive function and false-belief understanding in preschool children: Two tasks are harder than one. *Journal of Cognition and Development*, 4(3), 275-298.

Happé, F. G. E. (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

Happé, F., & Frith, U. (1996). Theory of mind and social impairment in children with conduct disorder. *British Journal of Developmental Psychology*, 14(4), 385–398. https://doi.org/10.1111/j.2044-835x.1996.tb00713.x

Harris, P. L. (2006). *Social Cognition*. In D. K. and R. S. W. Damon, R.M. Lerner (Ed.), Handbook of Child Psychology. Hoboken, NJ, USA: John Wiley & Sons, Inc. https://doi.org/10.1002/9780470147658.chpsy0219

Heyes, C. (2014). Submentalizing: I am not really reading your mind. *Perspectives on Psychological Science*, 9, 131–143.

Horn, J. L. (1994). Theory of fluid and crystallized intelligence. *Encyclopedia of Human Intelligence*, 1, 443-451.

Hughes, C., & Leekam, S. (2004). What are the links between theory of mind and social relations? Review, reflections and new directions for studies of typical and atypical development. *Social Development*, 13(4), 590–619. https://doi.org/10.1111/j.1467-9507.2004.00285.x

Hughes, C., & Russell, J. (1993). Autistic children's difficulty with mental disengagement from an object: Its implications for theories of autism. *Developmental Psychology*. https://doi.org/10.1037//0012-1649.29.3.498

Hutchison, K. A. (2007). Attentional control and the relatedness proportion effect in semantic priming. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 33(4), 645-662.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning: with applications in R. Corrected edition*. New York: Springer.

Javor, R. (2017). Bilingualism, Theory of Mind and Perspective-Taking: The Effect of Early Bilingual Exposure. *Psychology and Behavioral Sciences*, 5(6), 143.

Kahneman, D. (2011*). Thinking, fast and slow*. Macmillan.

Kan, K. J., Kievit, R. A., Dolan, C., & van der Maas, H. (2011). On the interpretation of the CHC factor Gc. *Intelligence*, 39(5), 292-302.

Kane, M. J., Bleckley, M. K., Conway, A. R., & Engle, R. W. (2001). A controlled-attention view of working-memory capacity. *Journal of Experimental Psychology: General,* 130, 169-183.

Keysar, B. (1994). The illusory transparency of intention: Linguistic perspective taking in text. *Cognitive psychology*, 26, 165-165.

Keysar, B., Lin, S., & Barr, D. J. (2003). Limits on theory of mind use in adults. *Cognition*, 89(1), 25–41. https://doi.org/10.1016/S0010-0277(03)00064-7

Kline, R. B. (2015). *Principles and practice of structural equation modeling*. Guilford publications.

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

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

Kovacs, K., & Conway, A. R. (2016). Process overlap theory: A unified account of the general factor of intelligence. *Psychological Inquiry*, 27(3), 151-177.

Lanza, E. (1992). Can bilingual two-year-olds code-switch? *Journal of Child Language*, 19(3), 633–658.

Lawson, J., Baron-Cohen S., and Wheelwright, S., (2004). Empathising and systemising in adults with and without Asperger Syndrome. *Journal of Autism and Developmental Disorders*, 34, 301-310

Lee, K., Olson, D. R., & Torrance, N. (1999). Chinese children's understanding of false beliefs: The role of language. *Journal of Child Language*, 26(1), 1–21. https://doi.org/10.1017/S0305000998003626

Leekam, S. R., & Prior, M. (1994). Can Autistic Children Distinguish Lies from Jokes? A Second Look at Second-order Belief Attribution. *Journal of Child Psychology and Psychiatry*, 35(5), 901–915. https://doi.org/10.1111/j.1469-7610.1994.tb02301.x

Legg, E. W., Olivier, L., Samuel, S., Lurz, R., & Clayton, N. S. (2017). Error rate on the director's task is influenced by the need to take another's perspective but not the type of perspective. *Royal Society Open Science*, 4(8), 170284.

Leslie, A. (1994*). ToMM, ToBy, and Agency: Core architecture and domain specificity*. In L. C. John Tooby, Alan M. Leslie, Dan Sperber, Alfonso Caramazza, Argye E. Hillis, Elwyn C. Leek, Michele Miozzo (Ed.), Mapping the Mind: Domain Specificity in Cognition and Culture (119–148).

Leslie, A. M., & Polizzi, P. (1998). Inhibitory processing in the false belief task: Two conjectures. *Developmental Science*, 1(2), 247–253. https://doi.org/10.1111/1467-7687.00038

Leslie, A. M., German, T. P., & Polizzi, P. (2005). Belief-desire reasoning as a process of selection. *Cognitive Psychology*, 50(1), 45–85. https://doi.org/10.1016/j.cogpsych.2004.06.002

Marian, V., Blumenfeld, H. K., & Kaushanskaya, M. (2007). The Language Experience and Proficiency Questionnaire (LEAP-Q): Assessing language profiles in bilinguals and multilinguals. *Journal of Speech, Language, and Hearing Research.*

Masangkay, Z. S., McCluskey, K. A., McIntyre, C. W., Sims-Knight, J., Vaughn, B. E., & Flavell, J. H. (1974). The Early Development of Inferences about the Visual Percepts of Others. *Child Development*, 45(2), 357–366.

McGrew, K. S. (2009). CHC theory and the human cognitive abilities project: Standing on the shoulders of the giants of psychometric intelligence research. *Intelligence*. 37(1), 1-10.

Miller, S. A. (2010). Social-Cognitive Development in Early Childhood. *Encyclopedia on Early Childhood Development.*

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

Mitchell, P., Robinson, E. J., Isaacs, J. E., & Nye, R. M. (1996). Contamination in reasoning about false belief: An instance of realist bias in adults but not children. *Cognition*, 59(1), 1-21.

Moll, H., & Tomasello, M. (2006). Level I perspective-taking at 24 months of age. *British Journal of Developmental Psychology*, 24(3), 603–613. https://doi.org/10.1348/026151005X55370

Moses, L. J. (2001). Executive accounts of theory-of-mind development. *Child Development*, 72(3), 688–690. https://doi.org/10.1111/1467-8624.00306

Naito, M., Komatsu, S., & Fuke, T. (1994). Normal and autistic children's understanding of their own and others' false belief: A study from Japan. *British Journal of Developmental Psychology*, 12(3), 403–416. https://doi.org/10.1111/j.2044-835x.1994.tb00643.x

Navarro, E. & Conway. Adult Bilinguals Outperform Monolinguals in Theory of Mind. Unpublished Manuscript.

Navarro, E., Macnamara, B. N., Glucksberg, S., & Conway, A. R. (2020). What Influences Successful Communication? An Examination of Cognitive Load and Individual Differences. *Discourse Processes*, 57(10), 880-899.

Nickerson, R. S. (1999). How we know - And sometimes misjudge - What others know: Imputing one's own knowledge to others. *Psychological Bulletin*, 125(6), 737–759. https://doi.org/10.1037/0033-2909.125.6.737

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, 818–823. doi:10.1037/ abn0000182

Obhi, S. (2012). The amazing capacity to read intentions from movement kinematics. *Frontiers in Human Neuroscience*, 6, Article 162. doi:10.3389/fnhum.2012.00162

Pennycook, G., Fugelsang, J. A., & Koehler, D. J. (2015). What makes us think? A three-stage dual-process model of analytic engagement. *Cognitive psychology*, 80, 34-72.

Perner, J. (1991). *Learning, development, and conceptual change. Understanding the representational mind*. The MIT Press.

Perner, J., Lang, B., & Kloo, D. (2002). Theory of mind and self-control: More than a common problem of inhibition. *Child Development*, 73(3), 752–767. https://doi.org/10.1111/1467-8624.00436

Piaget, J., & Inhelder, B. (1956). *The child's concept of space*. Routledge & Paul.

Poletti, M. et al. (2012) Cognitive and affective Theory of Mind in neurodegenerative diseases: neuropsychological, neuroanatomical and neurochemical levels. *Neurosci. Biobehav*. Rev. 36, 2147–2164

Premack, D., & Woodruff, G. (1978). Chimpanzee theory of mind. *Behavioral and Brain Sciences*, 4(1978), 515–526.

Preston, S. D., & De Waal, F. B. (2002). Empathy: Its ultimate and proximate bases. *Behavioral And Brain Sciences*, 25(1), 1-20.

Pylyshyn, Z. (1978). When is attribution of beliefs justified? *Behavioral and Brain Sciences*, 1(4), 592–593.

Quesque, F., & Rossetti, Y. (2020). What do theory-of-mind tasks actually measure? Theory and practice. *Perspectives on Psychological Science*, 15(2), 384-396.

Raven, J. C. (1938). *Progressive Matrices: Sets A, B, C, D, and E*. University Press, published by HK Lewis.

Royzman, E. B., Cassidy, K. W., & Baron, J. (2003). "I Know, You Know": Epistemic Egocentrism in Children and Adults. *Review of General Psychology*, 7(1), 38–65. https://doi.org/10.1037/1089-2680.7.1.38

Rubio-Fernández, P., & Glucksberg, S. (2012). Reasoning about other people's beliefs: Bilinguals have an advantage. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 38(1), 211.

Ruffman, T., Slade, L., & Crowe, E. (2002). The relation between children's and mothers' mental state language and theory-of-mind understanding. *Child Development*, 73(3), 734–751. https://doi.org/10.1111/1467-8624.00435

Russell, J. (1996). *Agency: Its role in mental development Hove*. Hove: Erlbaum (UK).

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

Saxe, R. R., Whitfield-Gabrieli, S., Scholz, J., & Pelphrey, K. A. (2009). Brain regions for perceiving and reasoning about other people in school-aged children. *Child development*, 80(4), 1197-1209.

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

Scholl, B. J., & Leslie, A. M. (2001). Minds, modules, and meta-analysis. *Child development*, 72(3), 696-701.

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

Sebanz, N., & Shiffrar, M. (2009). Detecting deception in a bluffing body: The role of expertise. *Psychonomic Bulletin & Review*, 16(1), 170-175.

Sodian, B. (1991). The development of deception in young children. *British Journal of Developmental Psychology*, 9(1), 173–188. https://doi.org/10.1111/j.2044-835x.1991.tb00869.x

Spearman, C. (1904). General Intelligence, Objectively Determined and Measured. *American Journal of Psychology*, 15, 201-93.

Spearman, C. E. (1927). *The abilities of man* (Vol. 89). New York: Macmillan.

Sperber, D., & Wilson, D. (1995). *Relevance: Communication and cognition*. Cambridge, MA: Harvard University Press.

Sperber, D., & Wilson, D. (2002). Pragmatics, modularity and mind-reading. *Mind & Language*, 17(1-2), 3-23.

Tabachnick, B. G., & Fidell, L. S. (2013). U*sing multivariate statistics: International edition.* Pearson2012.

Tardif, T., & Wellman, H. M. (2000). Acquisition of mental state language in Mandarin- and Cantonese-speaking children. *Developmental Psychology*, 36(1), 25–43.

Thurstone, L. L. (1938). *Primary mental abilities* (Vol. 119). Chicago: University of Chicago Press.

Van Overwalle, F., & Baetens, K. (2009). Understanding others' actions and goals by mirror and mentalizing systems: a meta-analysis. *Neuroimage*, 48(3), 564-584.

van Veluw, S. J., & Chance, S. A. (2014). Differentiating between self and others: an ALE meta-analysis of fMRI studies of self-recognition and theory of mind. *Brain imaging and behavior*, 8(1), 24-38.

Warnell, K. R., & Redcay, E. (2019). Minimal coherence among varied theory of mind measures in childhood and adulthood. *Cognition*, 191, 103997.

Wellman, H. M. (2018). Theory of mind: The state of the art. *European Journal of Developmental Psychology*, 15(6), 728-755.

Wellman, H. M., & Liu, D. (2004). *Scaling of Theory-of-Mind Tasks*, 75(2), 523–541.

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

Wimmer, H., & Perner, J. (1983). Beliefs about beliefs: Representation and constraircing function of wrong belifs in young children's understanding of deception. *Cognition*, (13), 103–128.

Wolf, E. J., Harrington, K. M., Clark, S. L., & Miller, M. W. (2013). Sample size requirements for structural equation models: An evaluation of power, bias, and solution propriety. *Educational and Psychological Measurement*, 73(6), 913-934.

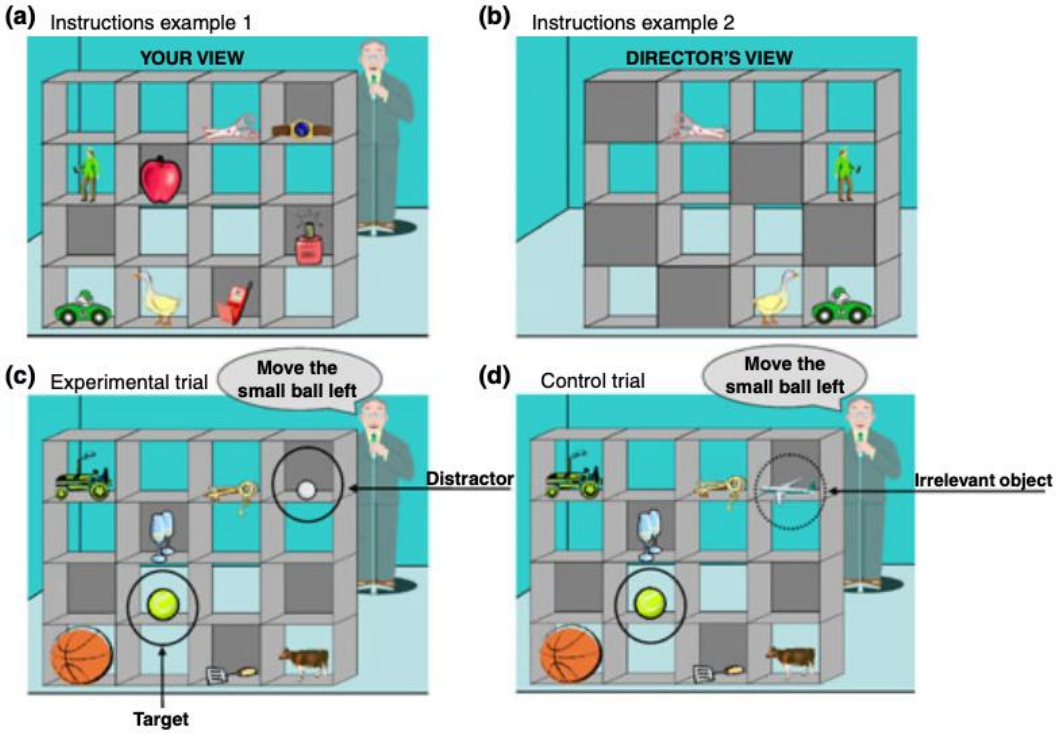
Zelazo, P., Muller, U., Frye, D., & Marcovitch, S. (2003). The Development of Executive Function in Early Childhood. The *Social History of the American Family: An Encyclopedia*, 68(3).

Zhou, M., & Shao, Y. (2014). A powerful test for multivariate normality. *Journal of Applied Statistics*, 41(2), 351-363.

Appendices

ix. Appendix A

Director task (Dumontheil, Apperly, & Blakemore, 2010). During the instructions phase, subjects were shown an example of their view (a) and the corresponding Director's view (b) for a given trial. During the experiment phase, subjects could encounter experimental trials (c) or control trials (d). Participants had to follow the oral instruction given by the Director. In experimental trials (c), the participant should move the target item (tennis ball) and ignore the distractor item (golf ball) if they took account of the Director's perspective. In control trials, an irrelevant object was shown instead of the distractor. Reprinted with Permission.



x. Appendix B

Sample Reading the Eyes in the Mind (RMET). Participants view the eye region of different faces and indicated the emotion that the eyes convey from four possible options.

jealous panicked

arrogant hateful

apologetic friendly

uneasy dispirited

joking flustered

desire convinced

xi. Appendix C

Sample section of the SSQ. Stories were adapted to American English and implemented in Qualtrics. Participants had to decide whether a statement was socially awkward.

STORY 1

(A) Christine had seen the advertisement on Tuesday.
'Shop assistant wanted for weekends and some evenings. Full training provided'.

So, the following morning was spent in town looking for a new dress to wear at the interview. After a quick coffee in 'Alfredo's', Christine made the decision to buy the dark blue one in 'Romanza'. It was a little overpriced but then again she could also wear it to her 50th birthday party in a few weeks. Also, as the shop assistant had said,

- "It seems to contrast with Madam's eyes so very well."
- "Do you really think so?" Christine had replied.
- "Oh yes," he added, "it highlights the blue most beautifully."

After returning to the shop and making her indulgent purchase, Christine walked to the bus station and caught the Number 7 bus home. As she got herself settled on the bus, the young conductor walked past.

- "Don't worry about showing me your O.A.P bus pass luv," said the conductor,
- "I can see you've got your hands full."

Q. Was anything said in the previous section that could have upset someone?

YES	☐	If yes, indicate what line it occurred in by filling in the circle
NO	☐	If no, please proceed to the next section

xii. Appendix D.

Counterbalancing analyses for Study 1 and Study 2.

Study 1

Additional analyses were conducted to examine whether there were differences task performance based on the order in which the tasks were presented. As mentioned in the Method section, the tasks were administered in one of three random orders. Order 1 was the following: Gf tasks, Gc tasks, ToM tasks. Order 2 was: ToM, Gf, Gc. Order 3 was: Gc, ToM, Gf. To simplify counterbalancing analyses, only one task per construct (i.e., the first task of each set of three tasks presented) was tested. Therefore, three three-way ANOVAs were conducted to examine differences in responses to each task for each order.

The first ANOVA compared accuracy in the Letter Series task of Gf across all three orders. There were no significant differences across orders ($F (2, 200) = 1.63$, $p = >.10$). The second ANOVA compared total responses to the Synonyms task of Gc across all three orders. There were no significant differences across orders ($F (2, 200) = .73$, $p = >.10$). The last ANOVA compared accuracy in experimental trials of the director task of ToM across order. There were no significant differences across orders ($F (2, 200) = .20$, $p = >.10$); the order in which participants completed the tasks did not affect performance in Study 1.

Study 2

Three additional analyses were conducted to examine whether there were differences in responses to the tasks based on the order in which the tasks were presented. As mentioned in the

Method section, the tasks were administered in one of four random orders. Order 1 was the following: director task, Simon task, verbal fluency, and metalinguistic tests. Order 2 was: Simon task, verbal fluency, metalinguistic tests, director task. Order 3 was: Verbal fluency, metalinguistic tests, director task, Simon task. Order 4 was: metalinguistic tests, director task, Simon task, verbal fluency. Four four-way ANOVA were conducted to examine differences in responses to each task for each order.

The first ANOVA compared accurate responses to the experimental trials of the director task across all four orders. The ANOVA reported a significant effect of Order (F (3, 311) = 5.71, p = .001). Post-hoc pairwise comparisons showed that the significant differences were between Order 4 and Order 2 (p = .005) and Order 4 and Order 3. (p = .001) indicating that participants performed more accurately when they were assigned to Order 4 compared to Orders 3 and 2. There were no other significant difference across orders for the director task.

The second ANOVA compared reaction times for incongruent trials of the Simon task across all four orders. The ANOVA reported a significant effect of Order (F (3, 152) = 3.52, p = .017). Post-hoc pairwise comparisons showed that the significant difference was between Order 2 and Order 1 (p = .016), indicating that participants performed more accurately when they were assigned to Order 2 compared to Orders 1. There were no other significant differences across orders for the Simon task.

The third and fourth ANOVA compared responses to the Graphophonemic and syntactic awareness tests, respectively. The ANOVA for the Graphophonemic test showed a significant effect of Order (F (3, 311) = 12.31, p = < .001). Post-hoc pairwise comparisons showed that the significant difference was between Orders 3 and Order 1 (p = <.001), Order 4 and Order 2 (p = .002) and Order 4 and Order 3 (p = <.001). The ANOVA for the syntactic awareness test showed

a significant effect of Order (F (3, 311) = 3.58, p = < .014). Post-hoc pairwise comparisons showed that the significant difference was between Order 4 and Order 2 (p = .008).

The last ANOVA compared responses to the verbal fluency task in English across all orders. The ANOVA showed a significant effect of Order (F (3, 311) = 6.77, p = < .001). Post-hoc pairwise comparisons showed that the significant difference was between Orders 2 and Order 1 (p = .014), Orders 3 and 1 (p = < .001), and Order 4 and 3 (p = .015).

Overall, the results of the counterbalancing analyses do not present a specific pattern of bias in one specific order, even though all ANOVA showed a significant difference in at least one order pairwise comparison. The Order that seems most problematic across all ANOVA is Order 4.

Tables

a) Table 1 - *Descriptive Statistics for Study 1.*

Variables		**Latent Construct**	***N***	***M***	***SD***	**Skew**	**Kurtosis**	**Range**	***α***
1. Ravens Progressive Matrices	RV	Fluid Reasoning (Gf)	203	9.28	3.84	-.89	-.78	0-18	.81
2. Letter Series	LS		203	6.77	2.59	.05	-.40	0-10	.77
3. Number Series	NS		203	9.40	3.07	-.26	-.61	0-15	.85
4. General Knowledge	GK	Crystallized intelligence (Gc)	203	7.09	2.26	-.99	.34	0-10	.79
5. Synonyms Task	SYN		203	5.99	2.52	-.63	-.33	0-10	.86
6. Antonyms Task	ANT		203	6.16	2.15	-.48	-.68	0-10	.85
7. Director Task	DT	Theory of Mind (ToM)	203	.59	.39	-.48	-1.61	0-1	.66
8. Reading the Eyes in the Mind	RME		203	30.14	4.02	-.92	1.12	0-36	.72
9. Short Stories Questionnaire	SSQ		203	10.15	3.33	-.20	-.24	0-20	.72

b) Table 2. *Correlations among variables.*

Variable	1	2	3	4	5	6	7	8
1. General Knowledge	-							
2. Synonyms	.50	-						
3. Antonyms	.43	.66	-					
4. Letter Series	.17	.19	.18	-				
5. Ravens	.52	.38	.44	.36	-			
6. Number series	.39	.33	.34	.49	.56	-		
7. SSQ	.18	.15	.15	.16	.33	.14	-	
8. RMET	.33	.32	.32	.30	.41	.39	.24	-
9. Director task	.16	.16	.16	.18	.32	.27	.18	.17

Note. All correlations were significant at $p = <.05$

c) Table 3. *Model Fit Indices for All CFA Models in Study 1.*

Fit Indices	χ^2	***df***	x^2/ ***df***	**CFI (TLI)**	**RMSEA**	**SRMR**
Recommended fit (Kline, 2015)			**≦2**	**≧.90**	**≦.08**	**.05 - .10**
Model 1: One predictor	164.19	28	5.86	.73 (.65)	.16	.149
Model 2: GF + TOM	95.51	32	2.98	.88 (.86)	.09	.157
Model 3: GF + GC + TOM	108.31	31	3.49	.85 (.83)	.11	.175
Network model	21.46	11	1.95	.98(.92)	.068	.033

Note. It is common that the Network model presents excellent fit, nevertheless, because the network model is an exploratory analysis, it is not possible to directly compare it to the CFAs.

d) Table 4. *Exploratory Factor Analysis Loadings for 3 factors. Tasks that loaded at >.30 were considered to load onto the same factor.*

Tasks	Factor 1 - Gc	Factor 2 - Gf	Factor 3 - ToM
1. Letter Series		.60	
2. Ravens		.55	
3. Number Series		.87	
4. Synonyms	.82		
5. Antonyms	.81		
6. General Knowledge	.49		
7. Director Task		.31	
8. SSQ			.84
9. RMET		.31	

e) Table 5a. *Demographic information for bilinguals and monolinguals.*

Group	Level of Education	Education Frequency	Reported Culture	Culture 1 Frequency	Culture 2 Frequency	Culture 3 Frequency
Bilingual	Doctorate	1	US-American	34	6	1
	Master	36	Hispanic	4	9	3
	Bachelor	33	Mexican	9	4	2
	Some college, no degree	5	Asian	1	3	1
	Associate degree	4	Black/African American	1	2	
	High school or equivalent	2	European	1		
			Catholic	3		7
			Native-American	1		
			White/ Caucasian	6	12	3
			Other		3	2
			Jewish		2	1
			Non-Hispanic		2	1
			Spanish		3	1
Monolingual	Master	9	US-American	81	4	
	Bachelor	34	Black/African American	5	2	
	Some college, no degree	28	Asian	3	2	
	Associate degree	17	White/ Caucasian	8	1	
	High school or equivalent	17	European	3	8	2
	Less than High School	1	Mexican-American	2		
			Catholic	1		
			Hispanic		1	1
			Other		2	2

*Note***.** Frequencies represent the number of participants who indicated a specific educational level or culture identification. Participants entered manually the culture or cultures they identified with. The maximum number of cultures reported were 3.

f) Table 5b. *Bilinguals' self-reported L1 (dominant) and L2 (nondominant).*

	L1	**L2**
English	81	2
Spanish	1	63

g) Table 5c. *Bilinguals' reported age of acquisition (AOA) of each language and years living in a country where each language is spoken.*

Mean AOA L1 (SD)	**Mean AOA L2 (SD)**	**Years in L1 country**	**Years in L2 country**
1.64 (1.78)	5.21 (4.3)	32.21 (17.62)	23.29 (18.66)

h) Table 5d. *Bilinguals' reported languages spoken across their lifespan.*

	0-2 years	**3-4 years**	**5-10 years**	**11-13 years**	**14-17 years**	**18-21 years**	**+22 years**
English	49	54	50	57	51	61	59
Spanish	17	19	32	23	35	29	28

i) Table 5e. *Bilinguals' average time using L1 and L2 by social group (%). Means (SD) are reported.*

	Friends	**Family**	**Coworkers**	**School**	**Religious events**	**Leisure**
L1	71.69 (21.07)	70.41(22.09)	69.59 (26.72)	70.79 (27.01)	63 (30.57)	71.09 (25.41)
L2	50.89 (27.56)	54.02 (23.03)	49.2 (30.66)	48.17 (30.12)	48.85 (30.45)	51.05 (30.94)

j) Table 6. *Descriptive statistics for measures in Study 2, reported separately for bilingual and monolingual subjects.*

	Variables	Construct	N	M	SD	Skew	Kurtosis	Range	α
B	**Graphophonemic task**	Metalinguistic Awareness	62	10.26	7.54	.57	-.80	0-30	.87
	Syntactic awareness		62	6.26	4.17	.92	.67	0-20	.87
	Verbal Fluency (English)	Verbal Ability	62	8.30	5.02	.52	-.20	0-20	1
	Simon Task (RT)	Executive Function	62	729.15	269.64	1.51	2.13	382-1683	.97
	Director Task	ToM	62	.16	.29	1.91	2.07	0-1	1
M	**Graphophonemic task**	Metalinguistic Awareness	92	15.32	6.81	-.52	-.37	0-30	.87
	Syntactic awareness		92	10.86	4.24	.21	-.54	2-23	.87
	Verbal Fluency (English)	Verbal Ability	92	12.95	3.91	-.22	.94	1-25	1
	Simon Task	Executive Function	92	575.40	154.85	1.98	5.05	373-1298	.97
	Director Task	ToM	92	.66	.33	-1.99	3.77	0-1	1

Note. Task reliability for the metalinguistic awareness tests was calculated by extracting Cronbach's alpha from the two tests. Task reliability for all other tasks was calculated by extracting Cronbach's alpha from the split in half dataset.

k) Table 7. *Correlation matrix among variables in Study 2.*

Variables	1	2	3	4
1. Verbal Fluency				
2. Director Task	.47			
3. Graphophonemic task	.60	.50		
4. Syntactic Awareness task	.52	.46	.54	
5. Simon task	-.47	-.38	-.30	-.43

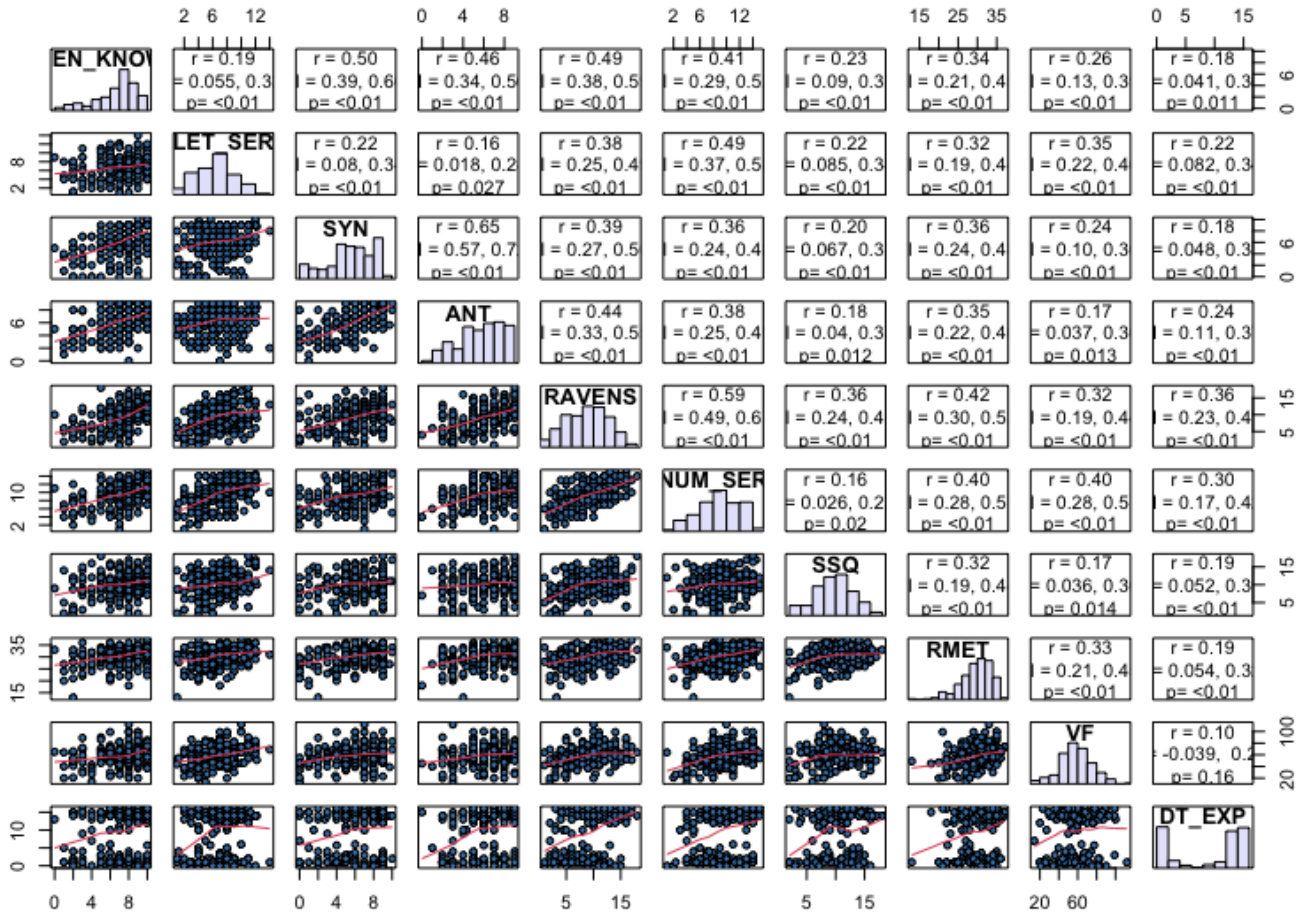
Note. All correlations were significant at $p < .05$.

l) Table 8. *Summary of Multiple Regression Analysis for Models 1-3, N=146).*

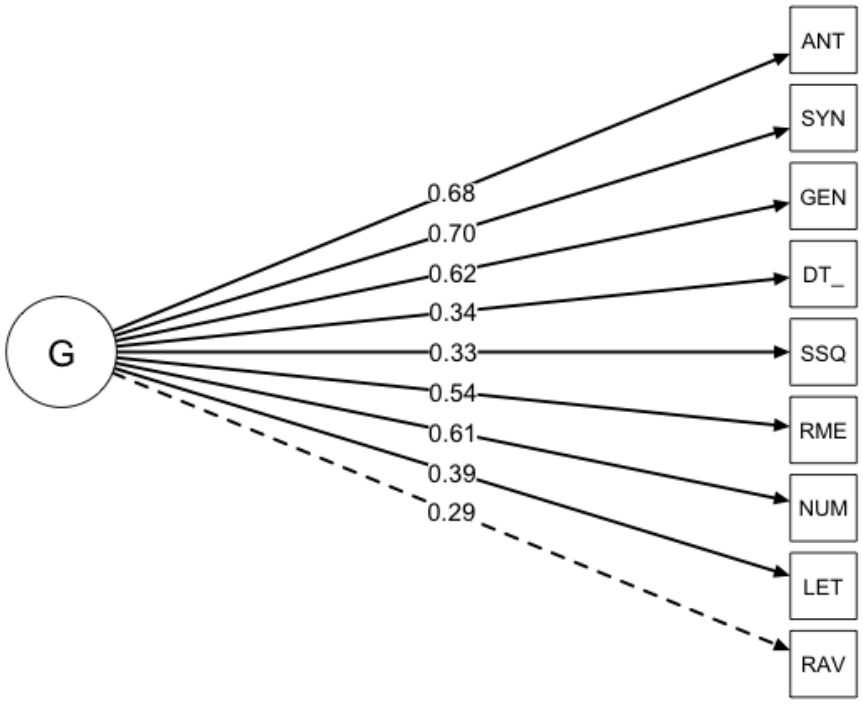
Model 1					
Variable	**B**	***SE***	***β***	**Odds ratio**	**P**
Simon task	-.003	.001	-2.28*	.996	.50
Metalinguistic Composite	.203	.051	4.03***	1.25	.55
Language Group	2.25	.528	4.27***	9.51	.90
Pseudo-R^2			-.708		
Model 2					
Variable	**B**	***SE***	***β***	**Odds ratio**	**P**
Simon task	-.002	.003	-1.27	.997	.499
Metalinguistic Composite	.21	.051	4.06***	1.23	.551
Language Group	3.68	2.15	1.71+	39.74	.975
Simon task *x* Group	-.002	.003	-.69	.997	.499
Pseudo-R^2			-.714		
Model 3					
Variable	**B**	***SE***	***β***	**Odds ratio**	**P**
Simon task	-.003	.001	-2.27*	.996	.499
Metalinguistic Composite	.205	.081	2.50*	1.23	.551
Language Group	2.28	1.38	1.66+	9.84	.908
Metalinguistic x Group	-.002	.103	-.027	.997	.499
Pseudo-R^2			-.708		

Note: *** <.001, ** <.01, *<.05, + <.1. P = probability.

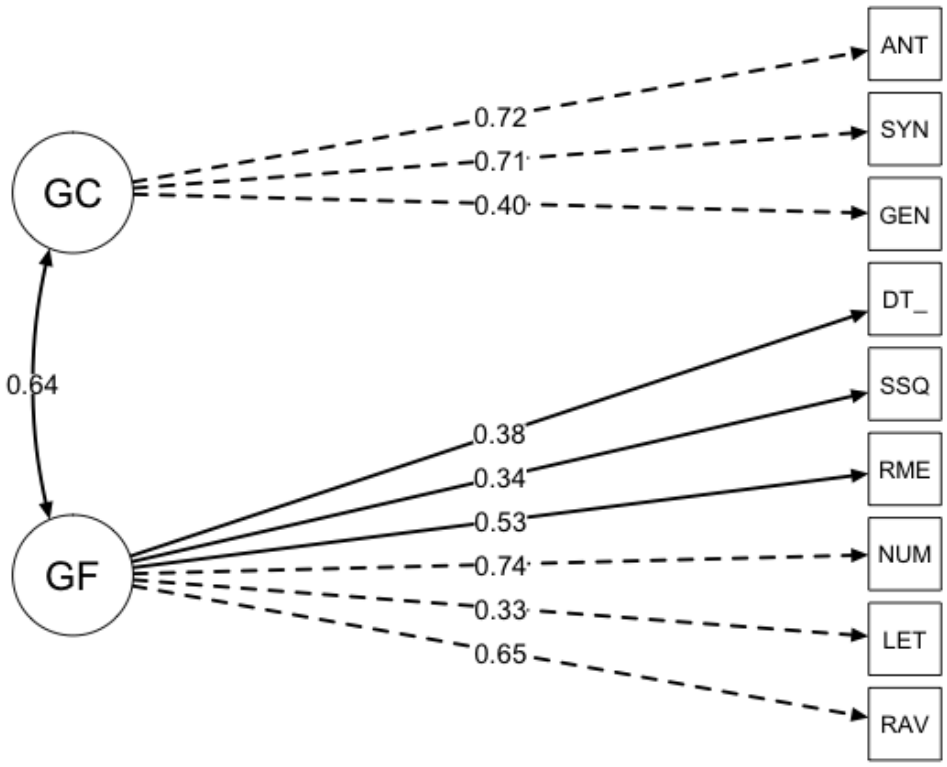
Figures



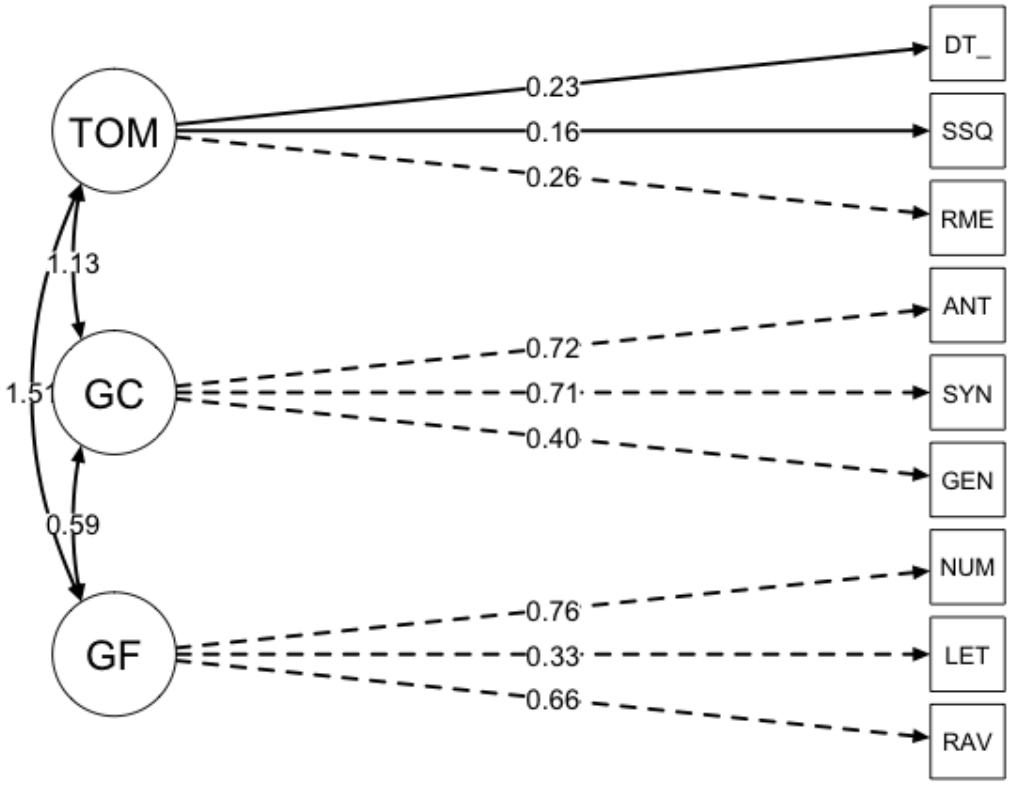
a) *Figure 1.* Plotted bivariate correlations and histograms for all tasks in Study 1.



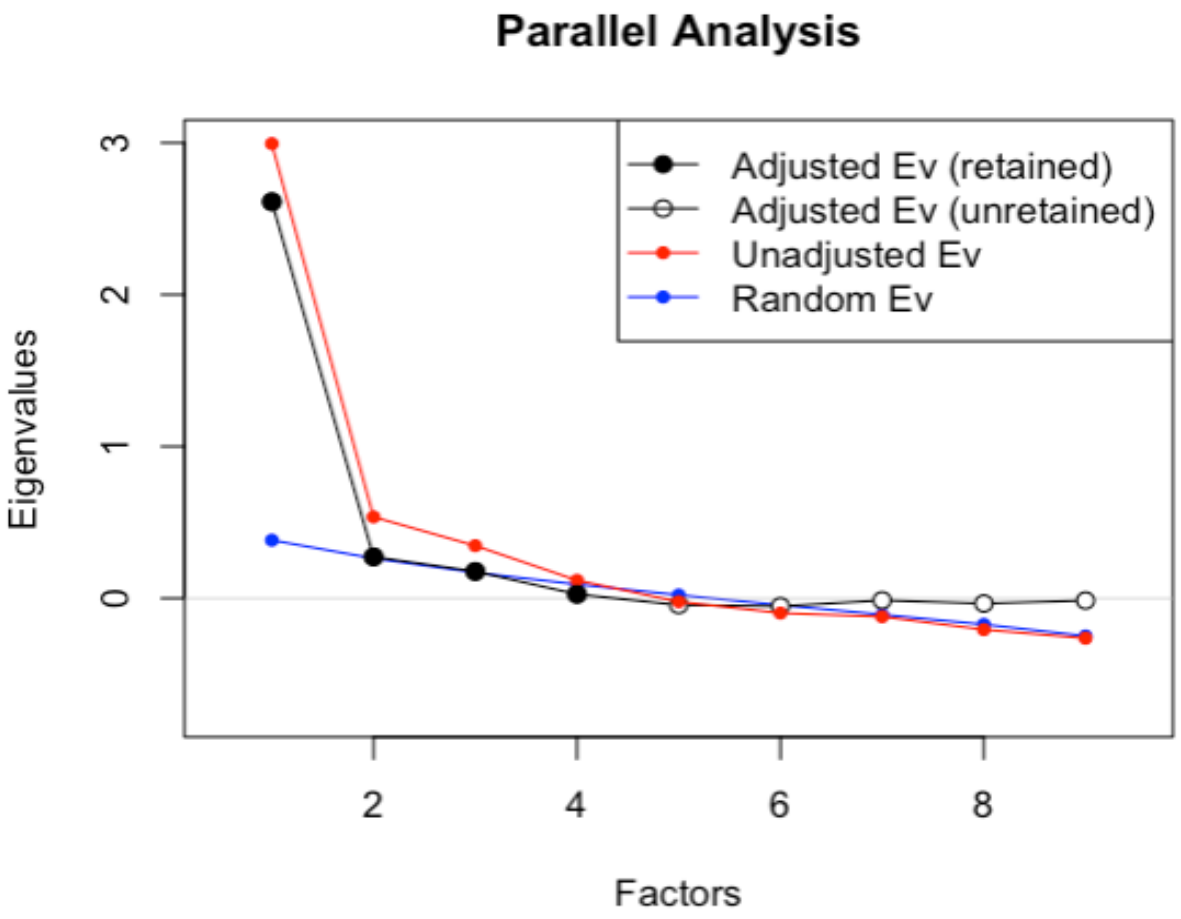
b) *Figure 2.* Model 1: One-factor model. Standardized factor loadings are presented. All loadings were within adequate range (>.30) with the exception of Raven's. ANT = Antonyms, SYN = Synonyms, GEN = General Knowledge, DT = Director task, SSQ = Short stories questionnaire, RME = Reading the eyes in the mind test, NUM = Number series, LET = Letter series, RAV = Raven's progressive matrices.



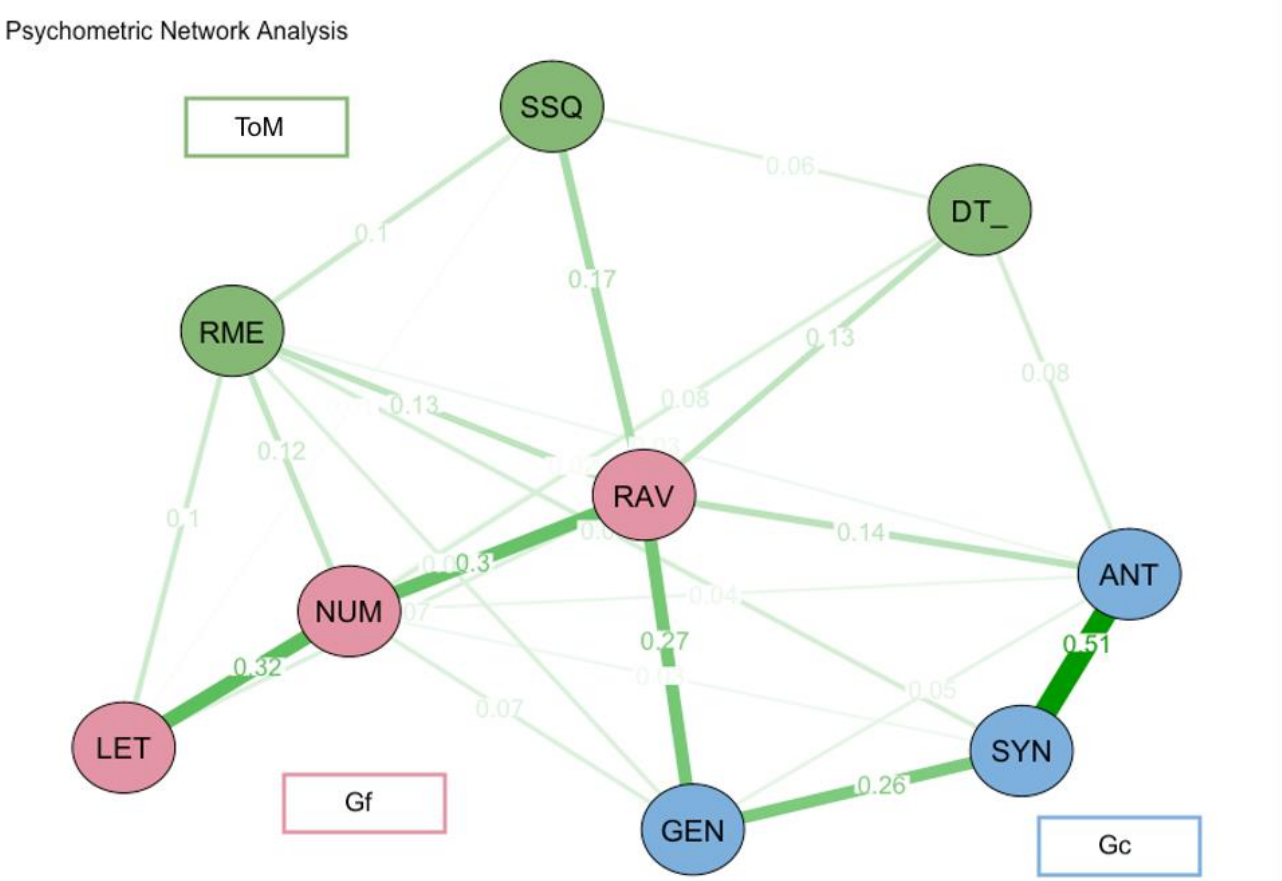
c) *Figure 3.* Model 2: Two-factor model. Standardized factor loadings are presented. All loadings were within adequate range (>.30). ANT = Antonyms, SYN = Synonyms, GEN = General Knowledge, DT = Director task, SSQ = Short stories questionnaire, RME = Reading the eyes in the mind test, NUM = Number series, LET = Letter series, RAV = Raven's progressive matrices.



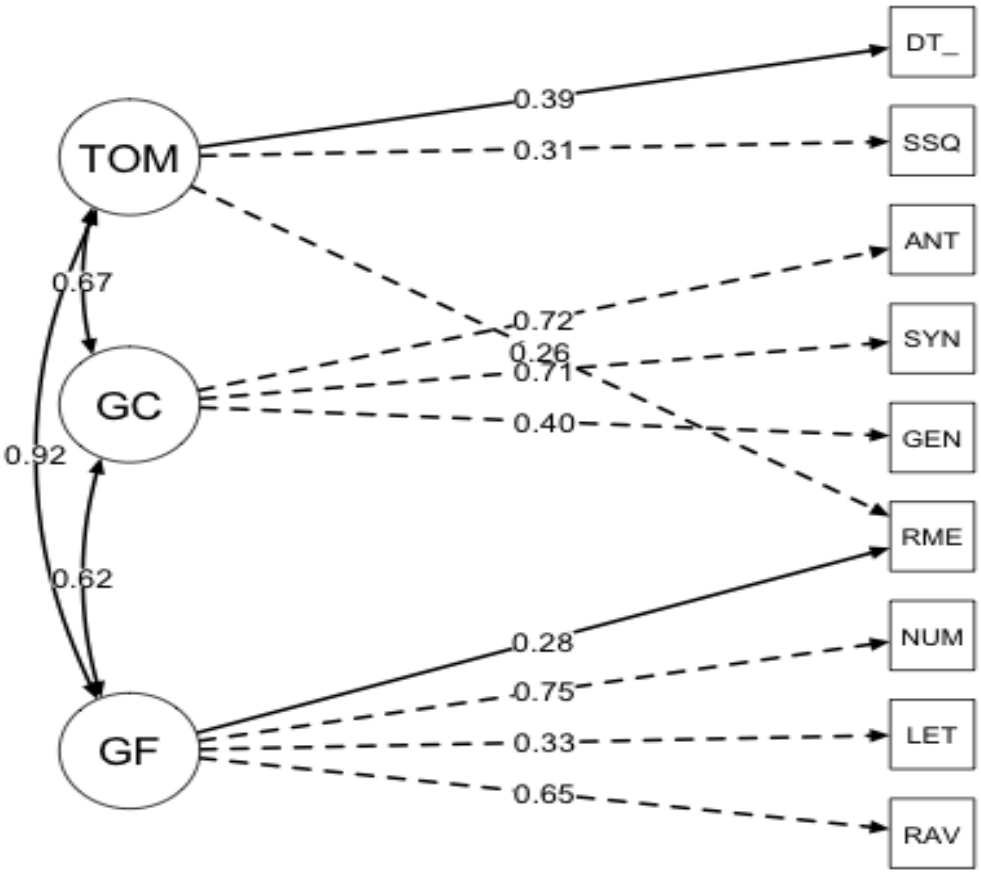
d) *Figure 4.* Model 3: Three-factor model. Standardized factor loadings are presented. Loadings were within adequate range for the Gc and Gf factors (>.30) but were low for the ToM factor (< .30). ANT = Antonyms, SYN = Synonyms, GEN = General Knowledge, DT = Director task, SSQ = Short stories questionnaire, RME = Reading the eyes in the mind test, NUM = Number series, LET = Letter series, RAV = Raven's progressive matrices.



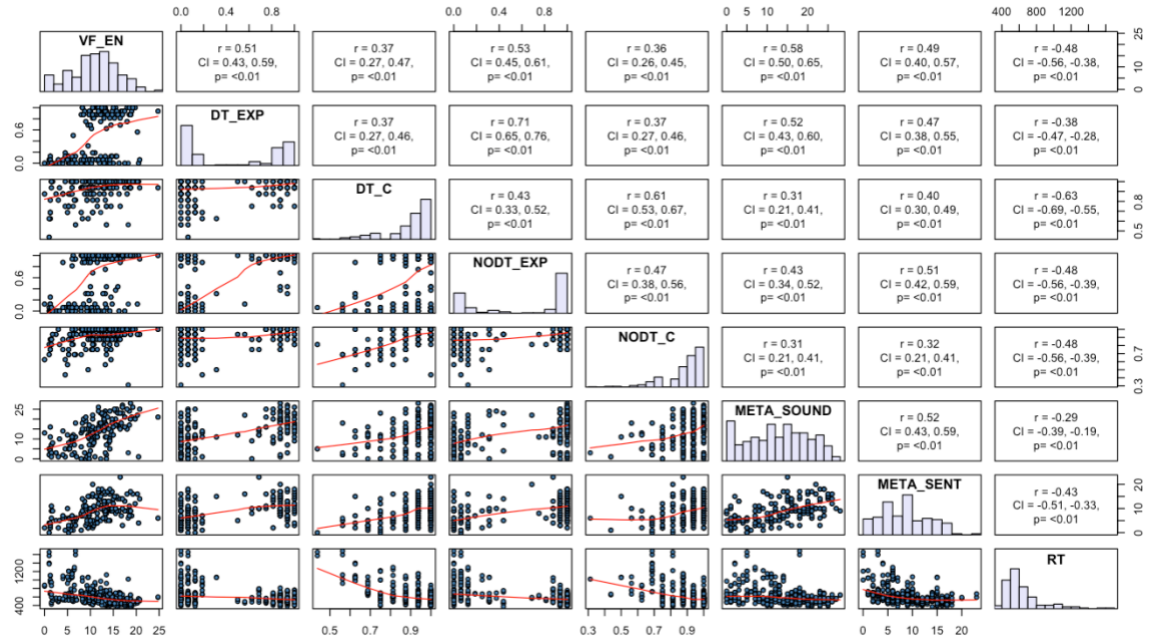
e) *Figure 5.* Parallel analysis. The black line represents the number of factors extracted from the dataset based on eigenvalues. The blue line represents the number of random factors retained from the random eigenvalues. The overlap of the blue and black line at factor 3 suggests that only factor 1 and 2 should be retained (Adjusted Eigenvalue 1 = 2.74, Adjusted Eigenvalue 2 = .29).



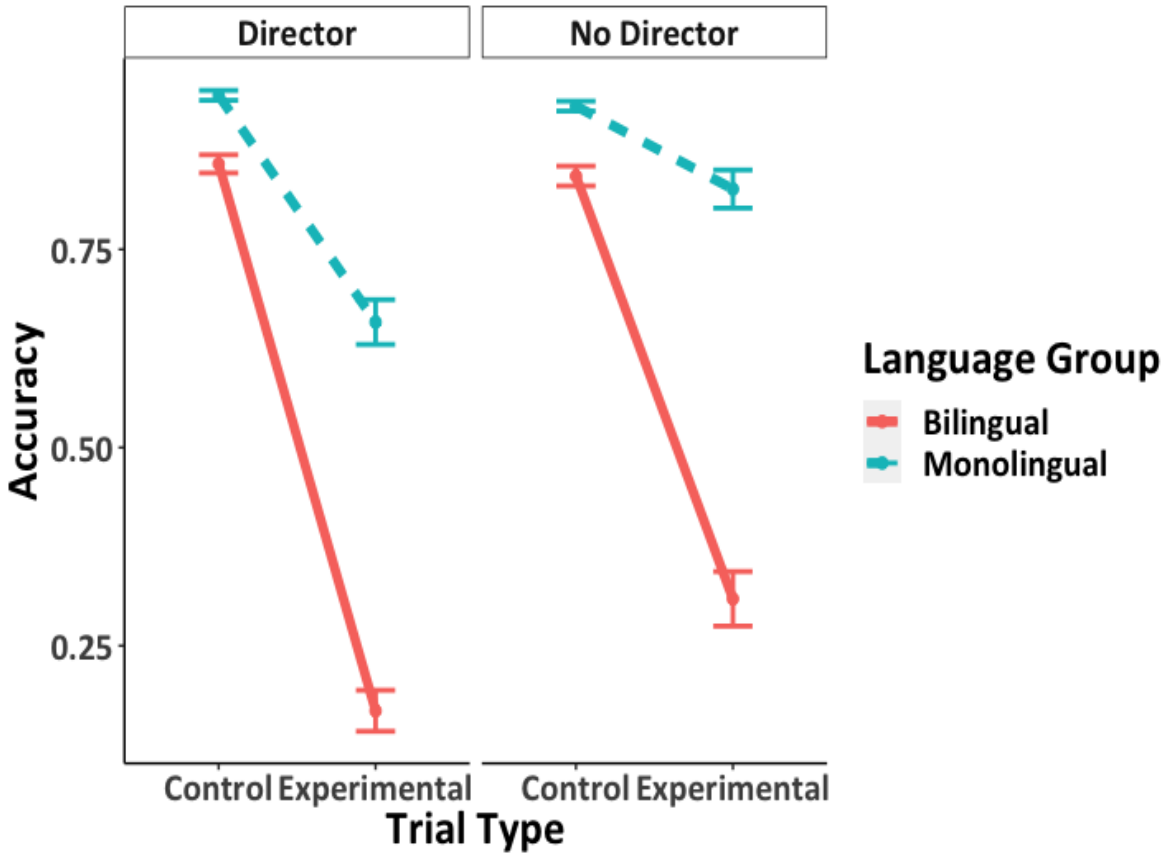
f) *Figure 6.* Network model. Nodes represent the tasks measured in the study. Edges represent the partial correlations among measures. Colors represent the theoretical construct they assess.



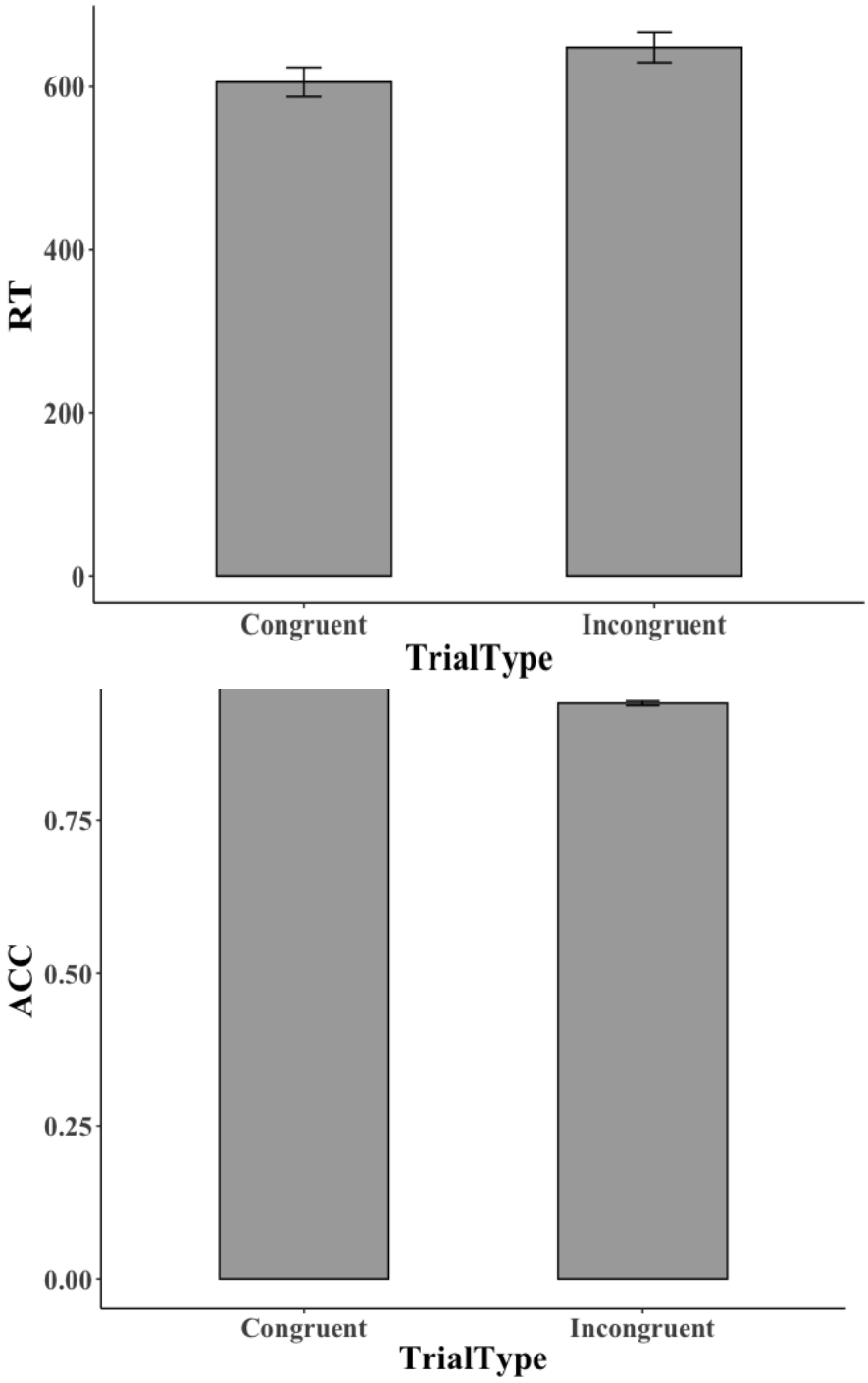
g) *Figure 7.* CFA model with modification indices (i.e., RMET is predicted by both ToM and Gf). ANT = Antonyms, SYN = Synonyms, GEN = General Knowledge, DT = Director task, SSQ = Short stories questionnaire, RME = Reading the eyes in the mind test, NUM = Number series, LET = Letter series, RAV = Raven's progressive matrices.



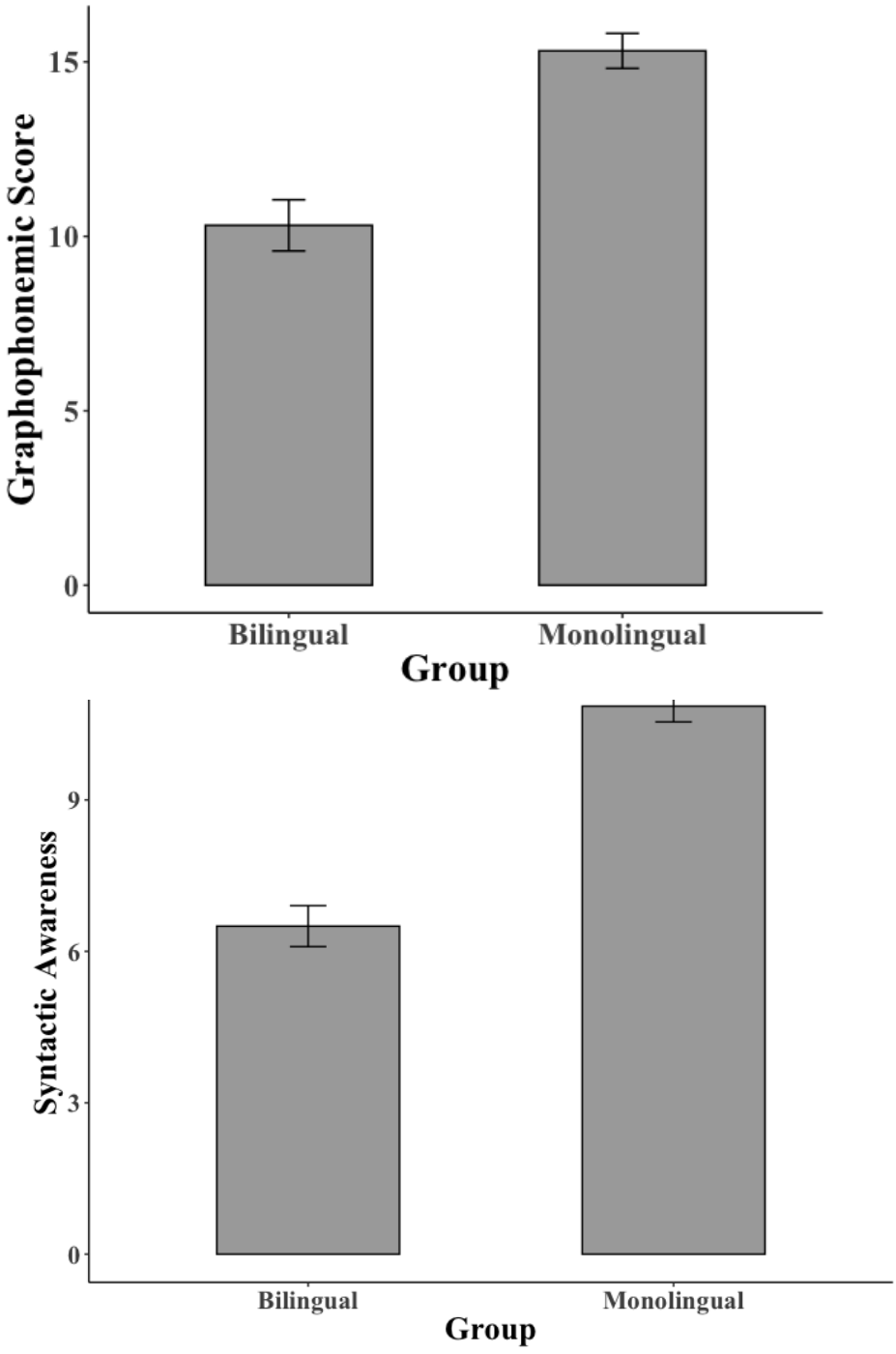
h) *Figure 8*. Plotted bivariate correlations and histograms for all tasks in Study 2.



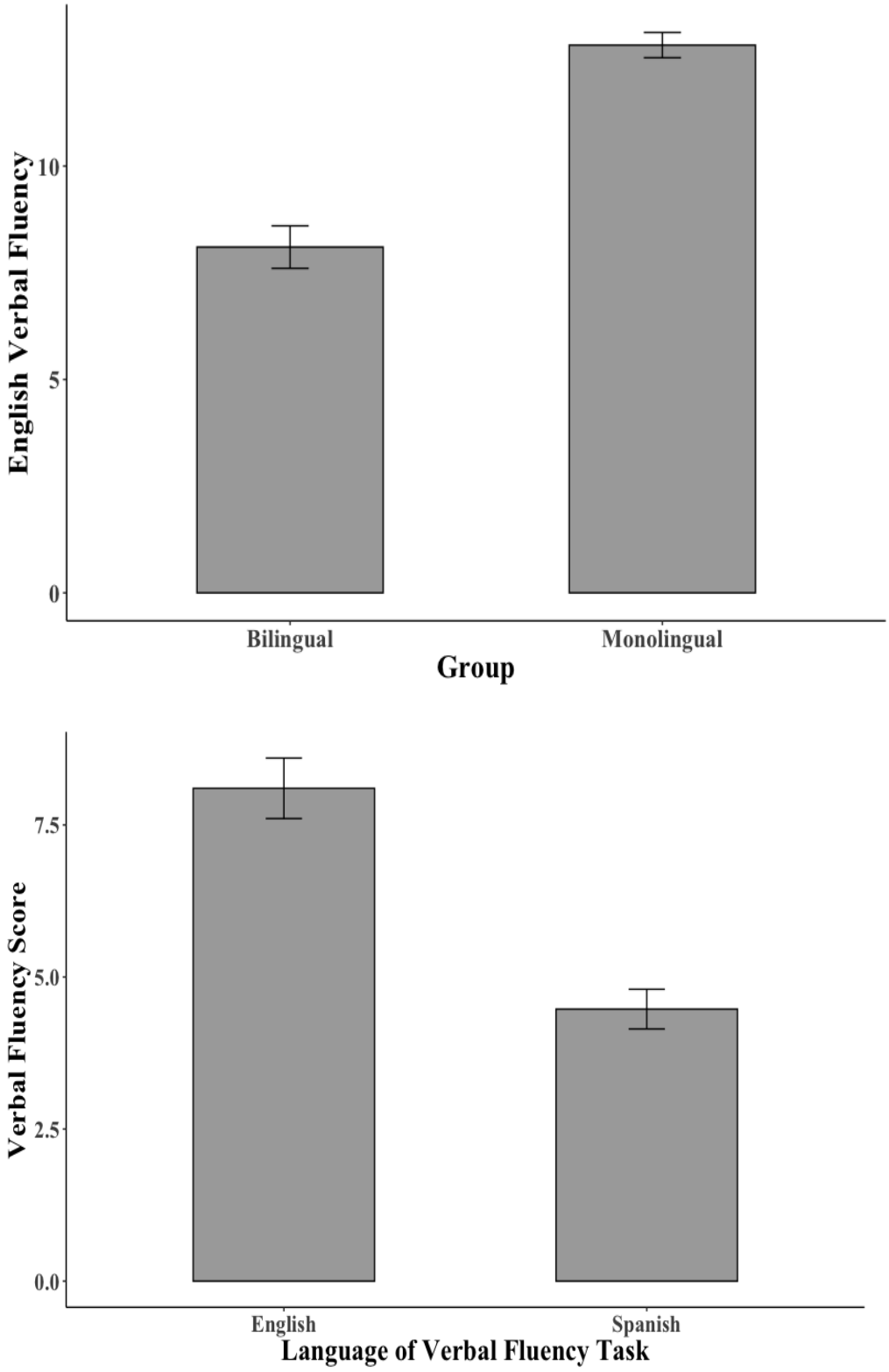
i) *Figure 9.* Response accuracy to the Director task by language group, condition, and trial type. The three-way interaction was not significant. There was a group by trial type interaction showing that monolinguals responded more accurately than bilinguals to experimental trials.



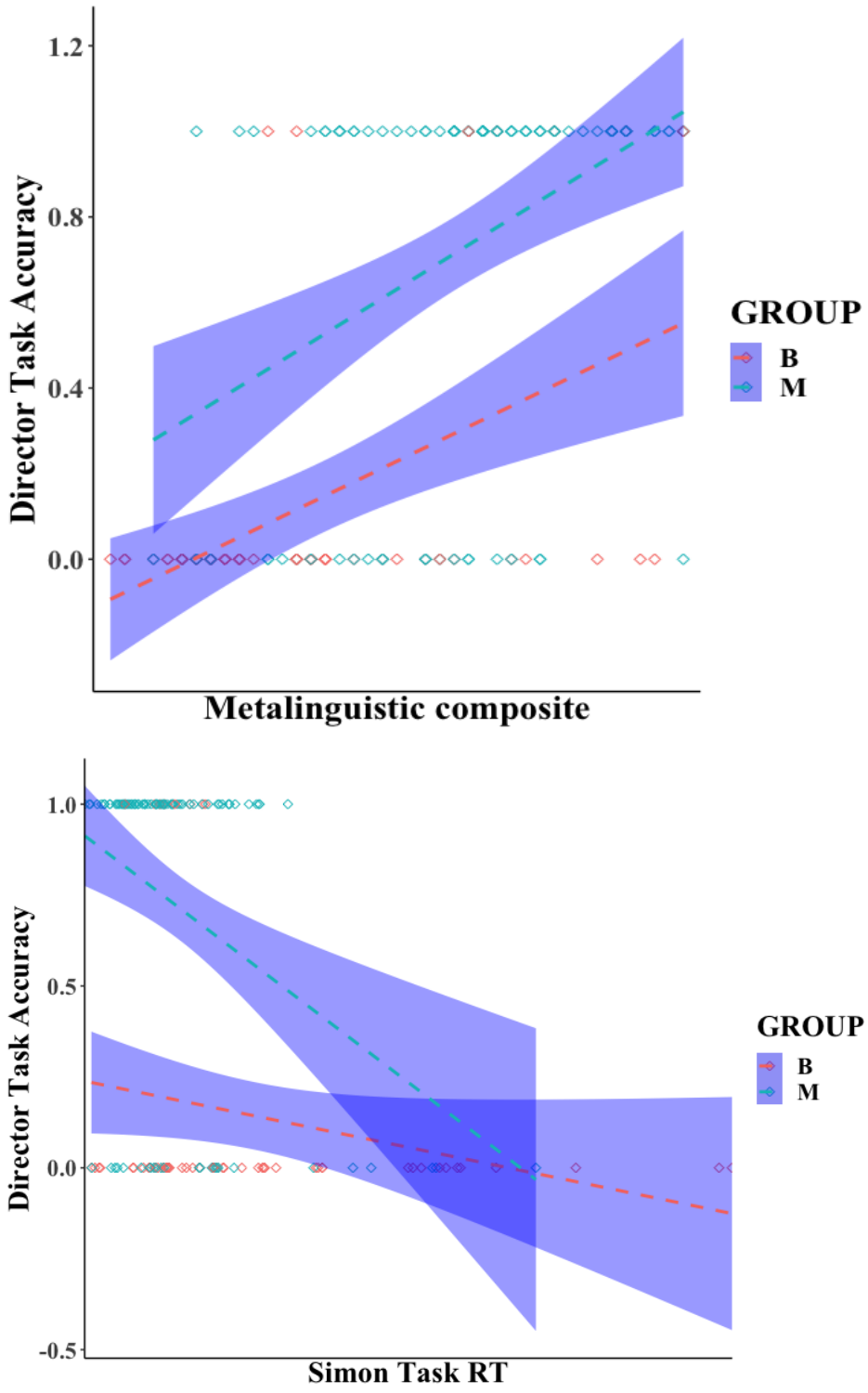
j) *Figure 10.* Response times (ms) to the Simon task by trial type and accuracy (proportion correct) to the Simon task by trial type.



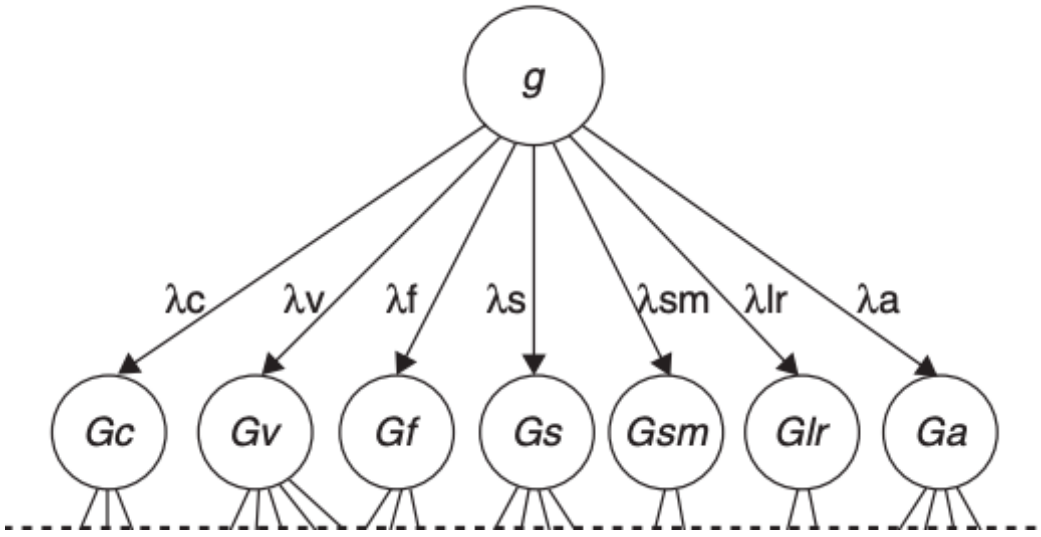
k) *Figure 11*. Total responses to the Graphophonemic test (i.e., correctly identifying words' phonemes) of Metalinguistic awareness by language group and total responses to the Syntactic Awareness test (grammatical sentences) of Metalinguistic awareness by language group.



l) *Figure 12*. Average Verbal Fluency score in English by group and Average Verbal Fluency score in Spanish compared to English for the bilingual group.



m) *Figure 13.* Models 1 and 2. Simon task and Metalinguistic awareness predicting the director task by group. Both predictor variables significantly predicted ToM but there was no moderation based on language group.



n) *Figure 14*. Cattel-Horn-Carrol (CHC) model of general intelligence and it's sub-abilities (McGrew, 2009).