SMALL GROUP THEORY OF MIND: BEYOND THE SINGLE AGENT FALSE BELIEF TASK

By

MICHELLE CHENG

A dissertation submitted to the

School of Graduate Studies

Rutgers, The State University of New Jersey

In partial fulfillment of the requirements

For the degree of

Doctor of Philosophy

Graduate Program in Psychology

Written under the direction of

Alan M. Leslie

And approved by

____________________________________

____________________________________

____________________________________

New Brunswick, New Jersey

January 2018
Our social world is complex and often includes multiple individuals with distinct thoughts or beliefs. Can preschoolers use their theory of mind to track the mental states and predict the actions of multiple individuals taking part in the same scenario? Theory of mind ability is traditionally tested in preschoolers using a single-agent false-belief task. In the main task, the Sally-and-Anne task, four-year-olds typically “pass” by predicting that an agent will search in accordance with her false belief, while three-year-olds typically “fail” by predicting that an agent will search in accordance to the child’s own belief. Some argue that three-year-olds’ failures are due to a conceptual deficit while others reason that failures are due to processing (e.g., executive function) demands such as working memory (WM) and inhibition. This suggestion of WM limitations in preschool theory of mind has never been directly tested, so we do not know what the WM limit for theory of mind may be nor whether it plays a role in limiting three-year-olds’ performance. The dissertation summarizes a series of seven studies designed to test preschoolers’ multiple-agent theory of mind on four WM levels: two, three, four, and five
agents each with distinct false beliefs combined with two levels of inhibition: low and high. The traditional single agent false belief task is tested with a binary response, pass/fail. This all but precludes the study of error patterns. Multiple agent false belief tasks, on the other hand, can for the first time provide data on how children fail when they fail, that is, provide for error analysis. We calculate the number of possible responses in a false belief task by raising the number of beliefs to the power of the number of agents, $B^A$. For example, with two agents, there are three possible beliefs (two false and one true) to the power of two agents (Sally and Anne), $B^A=3^2$. Of these nine responses, only one is ever fully correct and the remaining are errors of varying sorts. Although the number of responses differs across seven experiments, the two most common responses for both age groups were attributing the correct false belief to all of the agents and attributing the child’s own (true) belief to all of the agents. Inhibitory demand played the largest role in determining preschoolers’ performance across age groups: three-year-olds’ performed poorly in all of the high-demand tasks. In the low demand tasks, by contrast, differences between threes and fours were diminished or were absent. WM appeared to have an independent and incremental effect: three-year-olds’ performance appeared to drop off only when they were required to track four agents with distinct false beliefs while four-year-olds’ performance suffered with five agents. In both cases, evidence for a drop off appeared only in shifting error patterns rather than in their fully correct responding. These results strongly establish that WM capacity does not limit three-year-olds’ performance in the single-agent false-belief task. In sum, preschoolers are capable of tracking from three to five agents each with distinct false beliefs and acting within the same scenario. For every additional agent, the cognitive load increases only
slightly in comparison to the inhibitory demand of the standard Sally and Anne task. Theory of mind may have evolved to permit young children to track individuals in a small family-sized group.
Acknowledgements

First, I would like to thank the preschools and families who participated in the current series of studies. Their willingness to participate and to contribute to science is laudable and without them, this work would not have been possible. In addition, all of the undergraduate members of the Cognitive Development Lab have my gratitude for spending countless hours recruiting and coding subjects for my studies. Second, I would like to thank the graduate students who played an integral role in my graduate career. Special thanks to Dr. Lu Wang (a former Ph.D. student in the Cognitive Development Lab), who inspired the idea for the dissertation, Nic Kleene, who revised earlier versions of this dissertation, and Yelda Semizer, who joined me in researching into the wee hours of the night as well as weekends. Third, I would like to thank my brother, Brian Cheng, who created all of the stimuli that I have used in this dissertation as well as in other experiments that I ran throughout my graduate career. Fourth, I would like to thank all of those whom I have received mentorship from. Their guidance and feedback over the years have influenced me greatly. I owe many thanks to Dr. Pernille Hemmer, Dr. Judith Hudson, and Dr. Renee Baillargeon who served on my thesis committee. Their feedback improved this thesis tremendously. I would also like to thank Dr. Eileen Kowler, who piqued my interest in other areas of psychology that I would not have considered previously. I would like to thank Dr. Alan M. Leslie without whom none of this work would have been possible. He was a constant reminder of maintaining an appropriate work-life balance; he has inspired me with his enthusiasm for the field and attention to scientific rigor. This research was supported by NSF grant BCS-0922184.
# Table of Contents

Abstract of the Dissertation.................................................................ii
Acknowledgements....................................................................................v
Table of Contents....................................................................................vi
List of Tables............................................................................................ix
List of Illustrations..................................................................................x

1.0 Introduction.........................................................................................1
1.1 Executive function’s influence on theory of mind performance..............3
1.2 Theory of mind and working memory: “Holding in mind”......................5
1.3 Theory of mind and inhibitory control: Inhibiting a salient response.........8
1.4 Selecting between performance accounts.............................................14
1.5 Two explanatory accounts for how executive function affects theory of mind development.................................................................15
1.6 Limitations of the current theory of mind tasks....................................17
1.7 The current studies: Examining preschoolers’ ability to track multiple agents........18

2.0 Aims of Experiments 1-2......................................................................20
2.1 Experiment 1 (Double Agent High Demand)......................................20
2.1.1 Experiment 1 Method......................................................................20
2.1.2 Experiment 1 Results......................................................................24
2.1.3 Experiment 1 Discussion.................................................................28
2.2 Experiment 2 (Double Agent Low Demand)......................................30
2.2.1 Experiment 2 Method......................................................................30
List of Tables

1. Chapter 5, Table 1: Three-year-olds’ Probabilities of Success: Predicted vs. Observed Proportions..................................................................................................................85
2. Chapter 5, Table 2: Four-year-olds’ Probabilities of Success: Predicted vs. Observed Proportions..................................................................................................................85
3. Chapter 5, Table 3: Three-year-olds’ Probabilities of Failure: Predicted vs. Observed Proportions..................................................................................................................87
4. Chapter 5, Table 4: Four-year-olds’ Probabilities of Failure: Predicted vs. Observed Proportions..................................................................................................................87
5. Chapter 5, Table 5: Estimated Probabilities for the Single Agent Task: Three-year-olds on the Low Demand Tasks.................................................................................................90
6. Chapter 5, Table 6: Estimated Probabilities for the Single Agent Task: Four-year-olds on the Low Demand Tasks.................................................................................................90
7. Chapter 5, Table 7: Estimated Probabilities for the Single Agent Task: Three-year-olds on the High Demand Tasks.................................................................................................90
8. Chapter 5, Table 8: Estimated Probabilities for Single Agent Task: Four-year-olds on the High Demand Tasks.................................................................................................90
9. Chapter 5, Table 9: Examples of Responses Coded as “True Belief” responses in the Low Demand Tasks................................................................................................................95
List of Illustrations

1. Chapter 2, Figure 1: Layout of Experiment 1……………………………………22
2. Chapter 2, Figure 2: Preschoolers’ individual responses for Experiment 1……25
3. Chapter 2, Figure 3: Preschoolers’ individual responses for Experiment 1,
   separated by question order ........................................................................25
4. Chapter 2, Figure 4: Proportion of binned responses for Experiment 1……….27
5. Chapter 2, Figure 5: Layout of Experiment 2………………………………….32
6. Chapter 2, Figure 6: Preschoolers’ individual responses for Experiment 2……34
7. Chapter 2, Figure 7: Preschoolers’ individual responses for Experiment 2,
   separated by question order.........................................................................34
8. Chapter 2, Figure 8: Proportion of binned responses for Experiment 2……….35
9. Chapter 3, Figure 9: Layout for Experiment 3………………………………….43
10. Chapter 3, Figure 10: Preschoolers’ individual responses for Experiment 3....45
11. Chapter 3, Figure 11: Preschoolers’ individual responses for Experiment 3,
    separated by question order.....................................................................45
12. Chapter 3, Figure 12: Proportion of binned responses for Experiment 3……..46
13. Chapter 3, Figure 13: Proportion of binned responses separated by accuracy of
    memory recall for Experiment 3.................................................................47
14. Chapter 3, Figure 14: Layout of Experiment 4………………………………….51
15. Chapter 3, Figure 15: Preschoolers’ individual responses for Experiment 4……53
16. Chapter 3, Figure 16: Proportion of binned responses for Experiment 4……..54
17. Chapter 3, Figure 17: Proportion of binned responses separated by accuracy of
   memory recall for Experiment 4.................................................................56
18. Chapter 4, Figure 18: Layout for Experiment 5.............................................62
19. Chapter 4, Figure 19: Preschoolers’ individual responses for Experiment 5……64
20. Chapter 4, Figure 20: Proportion of binned responses for Experiment 5........65
21. Chapter 4, Figure 21: Proportion of binned responses separated by accuracy of
   memory recall for Experiment 5.................................................................66
22. Chapter 4, Figure 22: Layout for Experiment 6.............................................69
23. Chapter 4, Figure 23: Preschoolers’ individual responses for Experiment
   6............................................................................................................71
24. Chapter 4, Figure 24: Proportion of binned responses for Experiment 6........72
25. Chapter 4, Figure 25: Proportion of binned responses separated by accuracy of
   memory recall for Experiment 6.................................................................73
26. Chapter 4, Figure 26: Layout for Experiment 7.............................................77
27. Chapter 4, Figure 27: Preschoolers’ individual responses for Experiment 7......78
28. Chapter 4, Figure 28: Proportion of binned responses for Experiment 7.........79
29. Chapter 4, Figure 29: Proportion of binned responses separated by accuracy of
   memory recall for Experiment 7.................................................................80
30. Chapter 5, Figure 30: Predicted probabilities of success vs. observed
   proportions...............................................................................................86
31. Chapter 5, Figure 31: Predicted probabilities of failure vs. observed
   proportions...............................................................................................88
32. Chapter 5, Figure 32: Proportion of three- and four-year-olds’ all true belief errors in Experiments 1-7……………………………………………………………….91

33. Chapter 5, Figure 33: Individual responses for Experiments 1-7………………..93

34. Chapter 5, Figure 34: Count of true belief responses for each category………...96
Chapter 1

1.0. Introduction

This dissertation investigates how working memory and inhibition affect preschoolers’ verbal false belief performance using a new task that tests their ability to track multiple agents with distinct false beliefs. More than thirty years of research on false belief reasoning has, to our knowledge, only ever examined reasoning about a single agent with a single false belief. Here we present the first data on preschoolers’ reasoning about two, three, four, and even five agents each of whom have distinct false beliefs.

To understand an individual’s behavior, people must take into consideration the individual’s mental representation of the world. Researchers refer to this ability as “theory of mind” and test for its presence using a false belief task (Premack & Woodruff, 1978). In the Sally Anne false belief task (Baron-Cohen, Leslie, & Frith, 1985; see also Wimmer & Perner, 1983), preschoolers observe that Sally places a marble in a basket and leaves. In Sally’s absence, Anne moves the marble to a box. The critical question for preschoolers is, “When Sally returns, where will she search for her marble?” In order to predict Sally’s action correctly, preschoolers must recognize that while they may believe the marble is in the box, Sally falsely believes that the marble is in the basket. Children who consider Sally’s false belief will correctly predict that she will search in the basket. When verbally probed, four-year-olds typically succeed in attributing a false belief to Sally by predicting that she will search in the basket while three-year-olds typically fail to attribute a false belief to Sally by predicting that she will search in the box (Liu, Wellman, Tardif, & Sabbagh, 2008; Wellman, Cross, & Watson, 2001).
Preschoolers’ consistent behavior in the false belief task has led some researchers to suggest that children younger than four years of age do not possess theory of the mind, or more specifically, the concept of belief (Gopnik, 1993; Perner, 1991; Wellman, 1990). However, an increasing body of evidence has shown that infants and toddlers younger than three years of age can demonstrate false belief reasoning when their spontaneous responses are measured. For example, a variety of tests measuring eye-gaze can elicit false belief reasoning. First, infants look longer when the agent does not act according to her false belief (Kovacs, Telgas, & Endress, 2010; Onishi & Baillargeon, 2005; Scott, & Baillargeon, 2009; Scott, Baillargeon, Song, & Leslie, 2010; Song & Baillargeon, 2008; Song, Onishi, Baillargeon, & Fisher, 2008; Surian, Caldi, & Sperber, 2007). Second, toddlers prefer to look at the image that depicts an agent acting according to her false belief (Scott, He, Baillargeon, & Cummins, 2012). Third, toddlers anticipate that an agent will search according to her false belief (Clements & Perner, 1994; Garnham, & Ruffman, 2001; Southgate, Senju, & Csibra, 2007; Suran & Geraci, 2011; Wang & Leslie, 2016). Further, tests eliciting spontaneous helping have also demonstrated early false belief reasoning. More specifically, toddlers spontaneously help a mistaken agent (Buttelmann, Carpenter, & Tomasello, 2009; Buttelmann, Over, Carpenter, & Tomasello, 2009) by retrieving an ambiguously labeled toy (Carpenter, Call, & Tomasello, 2002; Happe & Loth, 2002; Southgate, Chevallier, & Csibra, 2010) and by pointing to the correct location (Knudsen & Liszkowski, 2012a; 2012b; Rhodes & Brandone, 2014).

Baillargeon, Scott, and He (2010; Carruthers, 2013; Friedman & Leslie, 2005; Leslie, 2000b; Setoh, Scott, & Baillargeon, 2016; Wang & Leslie, 2016 for a similar argument) proposed that the differences in theory of mind performance are due to
processing demands. More specifically, they argue that the standard verbal false belief tasks require more processing resources, typically described as executive function, than spontaneous response tasks. Evidence in favor of the processing account stems from a breadth of research that has found that executive function such as working memory and inhibition is positively correlated with theory of mind reasoning (Carlson, Moses, & Breton, 2002; Hala, Hug, & Henderson, 2003; Sabbagh, Moses, & Shiverick, 2006; see Duh, et al., 2016 for a cross-cultural study). Although it has often been suggested that preschoolers may fail because of limited working memory capacity (Gordon & Olson, 1998, Keenan, Olson, & Marini, 1998; Olson & Kamawar, 1999), there have been no attempts to directly test this suggestion and further, no attempts to determine what the limit on working memory capacity for tracking multiple agents and their mental states is.

For the remaining sections of Chapter 1, we outline the current perspectives on the relationship between executive function and theory of mind and the associated correlational studies supporting these perspectives. Chapters 2-4 present a series of seven experiments, which aim to answer three questions. First, which executive function ability limits three-year-olds’ performance in the standard Sally Anne task, working memory (WM) or inhibitory demand (ID)? Second, what is the WM capacity limit for tracking multiple agents with distinct false beliefs? Third, because this work, for the first time in theory of mind research, makes possible an error analysis of children’s performance, what type of errors do preschoolers make in the false belief task? Chapter 5 compares performance across all the studies and asks what the pattern of success and the multiple patterns of failure reveal about theory of mind reasoning.

1.1 Executive function’s influence on theory of mind performance
Executive function (EF) refers to the set of abilities that demonstrates an individual’s cognitive flexibility, which includes the ability to shift between tasks, to plan future goals, to manipulate items in memory (WM), and to suppress prepotent responses (inhibition). The two EF components that researchers typically relate to theory of mind reasoning are the latter two skills, WM and inhibition. Studies have shown that: 1) both positively correlate with false belief performance (Carlson et al., 2002; Hala et al., 2003; see Duh et al., 2016 for a cross-cultural study) and 2) limitations in each can explain younger children’s systematic failures in the verbal false belief task. More specifically, when a child incorrectly points to the marble’s current location in the Sally Anne task, it is possible that the child is unable to track both Sally’s outdated representation and the current representation of the world simultaneously (i.e., a WM error; Davis & Pratt; Keenan, 1998; Keenan et al., 1998). It is also possible that the child is unable to overcome the prepotent response to attribute their own belief of the marble’s location (i.e., an inhibition error; Leslie & Polizzi, 1998; Leslie & Thaiss, 1992; Roth & Leslie, 1998).

In the following sections, we will first review how WM can influence theory of mind performance and the evidence that supports this relationship. Following the relationship between WM and theory of mind, we will review how inhibitory control can influence theory of mind performance and the evidence that supports this relationship. We will then summarize an argument stemming from pretense that discounts the WM account. In addition to the role that EF skills play in theory of mind performance, we will also discuss two theories that describe how EF influences theory of mind development. In the final section of this chapter, we will discuss how the current false belief tasks are
unable to determine neither which EF skill plays a role in theory of mind performance nor how EF plays a role in theory of mind development, and how the current studies will remedy this limitation.

1.2. Theory of mind and working memory: “Holding in mind”

According to Baddeley and Hitch (1974; see also Baddeley, 1981; Baddeley, 1998), WM is a cognitive process in which information can be encoded and retrieved for a short period. Its capacity is limited (Baddeley, 1986) and develops with age (Pasual-Leone, 1970). Indeed, a wealth of studies have shown that WM capacity increases during the preschool years (from three- to five-years-olds; Ewing-Cobbs, Prasad, Landry, Kramer, 2004; Hongwanishkul, Happaney, Lee, & Zalazo, 2005; Hughes, 1998b). The developmental increase of WM capacity in the preschool years has led Olson and others (Davis & Pratt, 1995; Keenan, 1998; Keenan et al., 1998) to propose that younger children’s WM capacity limits their performance in false belief tasks. They reason that WM is necessary to succeed in false belief tasks because children are required to simultaneously track and maintain two representations of the world: Sally’s outdated representation and the current representation (Davis & Pratt, 1995; Keenan, 1998; Keenan et al, 1998; McKinnon & Moscovitch, 2007; Olson, 1989). It is hypothesized that three-year-olds’ WM capacity is limited to one representation, and therefore, they fail false belief tasks because they can only hold one representation at a time. In other words, systematic failure is due to replacing Sally’s outdated representation (i.e., Sally’s false belief) with the current representation while systematic success is due to a larger WM capacity, i.e., a capacity of at least two representations.
Several studies support the relationship between WM capacity and theory of mind. These studies have used various methods to measure WM capacity and then correlated WM performance and false belief reasoning performance (Davis & Pratt, 1995; Gordon & Olson, 1998; Keenan et al., 1998). The first is Gordon and Olson’s (1998) experiment where researchers measured WM capacity by an individual’s performance on the “Counting and Labeling” task. This task required children to complete three different steps: 1) point and numerically label a set of objects, 2) point and categorically label the same objects, and finally 3) point and then both numerically and categorically label each object. Researchers then assessed children’s false belief reasoning by using the three types of verbal false belief tasks. Gordon and Olson (1998) found that children’s performance on the Counting and Labeling task positively correlated with their performance on all variations of the verbal false belief task, suggesting that children’s WM capacity could be limiting their false belief performance.

In another study examining the relationship between WM capacity and theory of mind performance, Keenan et al. (1998) measured children’s WM capacity in a Counting Span task. In this task, experimenters showed children character cards with various colored dots and instructed children to count the number of red dots and ignore the other colored dots. After the child successfully counted the targeted dots, the experimenter flipped the card, so the dots were not visible to the child. The experimenter then asked the child to recall the number of dots on the character’s card. If the child recalled the number of dots successfully, the experimenter added another character card. Children were then required to recall the number of dots for both character cards. This continued until children failed to recall the number of dots correctly for two consecutive trials. The
highest number of character cards successfully recalled was considered the child’s WM capacity. Researchers found that children’s WM capacity in the Counting Span task correlated with performance in two types of false belief tasks: the change location task (Sally Anne task) and the unexpected contents task. Finally, Davis and Pratt (1995) measured children’s WM capacity in a Forward Digit Span task, where children were instructed to recall the numbers in a digit string, and Backward Digit Span task, where children were instructed to recall a digit string in reverse order. Again, performance in both WM tasks correlated with performance in the performance in two of the three standard false belief tasks: the unexpected contents task and the appearance-reality task.

Three-year-olds’ limited WM capacity should also have implications in other tasks that require tracking more than one representation. If WM capacity limits three-year-olds to tracking only a single representation of the world at a time, then three-year-olds’ performance on other tasks that require dual representations, such as the false photographs task, should be similar to their false belief performance. In the false photographs task (Zaitchik, 1990; 1991), participants were presented with a miniature model room that included a target toy (a duck), a bed, and a chair. The room was arranged so that the duck was sitting on top of the bed and a chair was placed beside the bed. Participants were first allowed to examine the pre-arranged room. Once the participant was satisfied, the experimenter took a picture of the room. Participants watched as the photograph developed, and then the experimenter placed the photograph face down. The experimenter then moved the duck from the bed to the chair while narrating their actions aloud. Participants were then asked to recall the duck’s location in the photograph. Like the false belief task, there are two different representations of the
duck’s location. The first is the outdated representation (Sally’s false belief about the marble’s location in the false belief task and the duck’s location in the photograph in the false photographs task). The second is the current representation (the marble’s current location in the false belief task and the duck’s current location in the false photographs task). Thus, if three-year-olds are limited to track and maintain only one representation (like the Sally-Anne task), they should also perform poorly on the false photographs task by recalling the current representation. In contrast, four-year-olds, who have the capacity to retain at least two representations, should successfully recall the outdated representation. This is exactly what Zaitchik (1990; 1991) found. Like the false belief task, three-year-olds incorrectly responded with the current representation while the four-year-olds correctly responded with the outdated representation. However, this finding along with similar performance in the standard Sally-Anne task could also be explained by immature inhibitory control. In the following section, we define two subsets of inhibition and then describe its relation to false belief reasoning.

1.3. Theory of mind and inhibitory control: Inhibiting a salient response

Studies have found that inhibitory control develops in the preschool years (Gerstadt, Hong, & Diamond, 1994; Jerger, Martin, & Pirozzolo, 1988; Livesey & Morgan, 1991; Reed, Pien, & Rothbart, 1984). Researchers have characterized inhibition in two ways: “delay” inhibition and “conflict” inhibition. Delay inhibition is the ability to wait for an outcome. Researchers typically measure delay inhibition by administering tasks such as the gift delay task, the tower building task, (Kochanska, Murray, Jacques, Koenig, & Vandergeest, 1996) and the marshmallow task (Mischel, Ebbesen, & Raskoff Zeiss, 1972). In these tasks, children are required to inhibit the impulse to peek at a gift
that is being wrapped (gift delay task), inhibit the impulse to continue to build a tower and skip their partner’s turn (tower building task), and inhibit the impulse to take an immediate reward in favor of a more gratifying reward in the future (marshmallow task).

Conflict inhibition, on the other hand, is the ability to inhibit a salient response. Researchers commonly measure conflict inhibition by administering tasks such as the Day/Night task (Gerstadt et al., 1994) and the Bear/Dragon task (Kochanska et al., 1997; Reed, Pien, & Rothbart, 1984). The Day/Night task is a child-friendly Stroop task where children are required to inhibit salient word-meaning associations. For example, children were shown two sets of cards: “Moon” cards and “Sun” cards. The goal of the task was to state “Day” when shown a moon card despite having strong associations of nighttime with the moon and “Night” when shown a sun card despite having strong associations of daytime with the sun. To perform successfully in the Day/Night task, children must resolve the conflict between their semantic knowledge and the correct verbal response.

The Bear/Dragon task also examined conflict inhibition by using a modified Simon Says task. Children in this task must inhibit the prepotent response to imitate the actions that they observe. The goal of task was to imitate only the Bear’s actions and not the Dragon’s. To be successful in this task, children must resolve the conflict of imitating one of the puppet’s actions but not the other. Researchers have found that performance on tasks that require the latter subcategory of inhibition, conflict inhibition, correlates with performance on false belief tasks and can even predict future theory of mind performance (Baker, Friedman, & Leslie, 2010; Carlson & Moses, 2001; Carlson, Moses, & Hix, 1998; Hala et al., 2003; Hughes, 1998a, 1998b).
Why is inhibition, specifically conflict inhibition, required in a false belief task? Leslie and Thaiss (1992) proposed that suppressing a salient response is essential to be successful for both the false photographs task and the Sally Anne task. In the false photographs task, for example, children face selecting between two responses: children can choose to either respond with the outdated representation (correct) or respond with the current representation (incorrect). The most salient response in the false photographs task is the target’s present location because it is in the child’s current view, while the less obvious, albeit, correct answer is not visible to the child. They must inhibit the salient response to be able to select the correct response. Likewise, children face the same type of response selection in a false belief task.

Leslie and colleagues (Friedman & Leslie, 2005; Leslie & Polizzi, 1998; Roth & Leslie, 1998) have argued that belief attribution is a process in which individuals must first represent all plausible belief candidates and then execute an additional process that selects the most salient candidate. The most salient candidate in belief attribution tasks is initially the true belief because beliefs are typically true, and therefore, selecting the true belief is the “default” strategy when reasoning about others’ beliefs. Choosing the “default” candidate in a false belief task, however, will lead an individual to the incorrect belief attribution because the correct attribution for a false belief task is the false belief and not the true belief. According to Roth and Leslie (1998; see also, Friedman & Leslie, 2005; Leslie & Polizzi, 1998) to overcome the true belief default strategy, the belief-attributer must inhibit the true belief to reduce its saliency. With the true belief’s saliency reduced, the belief-attributer will then be able to select the false belief candidate. Depending on the individual’s level of inhibitory power, their inhibition will either: 1)
adequately reduce the saliency of the true belief candidate, which results in correctly selecting the false belief candidate or 2) inadequately reduce the saliency of the true belief candidate, which results in incorrectly selecting the true belief (Friedman & Leslie, 2005; Leslie & Polizzi, 1998). Thus, to pass a false belief task, children must apply sufficient inhibition to the true belief candidate to reduce its saliency, so they can correctly select the false belief candidate.

The selection process model offers a subtly different account of the characteristic false belief error from another inhibitory account called the “pull of the real” account (Carpenter et al., 2002; also called “the reality bias,” Mitchell, 1994; Russell, Mauthner, Sharpe, & Tidswell, 1991). The “pull of the real” refers to children’s tendency to use reality as a response (i.e., report where the object is currently located), implying that children are not reasoning about an individual’s mental state but only about the state of the physical world. In contrast, Leslie’s selection process model indicates that children who fail the Sally Anne task are representing the individual’s mental state, but are failing to select the correct belief candidate. Kikuno, Mitchell, and Ziegler’s (2007) theory of mind reaction time study provides evidence in favor for Leslie’s selection model. In this study, preschoolers’ reaction times for three questions in the false belief task, the memory question, the reality question, and the action prediction question, were recorded (Kikuno et al., 2007). It is noteworthy that the responses for the two comprehension questions: “Where did Sally put the marble in the beginning of the story?” (Memory question) and “Where is the marble right now?” (Reality question) and the two possible responses for the action prediction question: the false belief response and the true belief response are the same responses respectively. The former responses, however, do not require belief
reasoning while the latter does. Kikuno and colleagues (2007) found that preschoolers who incorrectly attribute a true belief took longer to respond the action prediction question than the reality question despite reporting the same verbal answer. This finding suggests that when children incorrectly attribute a true belief, they are not simply recalling the marble’s current location, but rather they are reasoning about an individual’s belief.

Unlike WM demand in false belief tasks, the influence of inhibition on false belief performance has been examined. When ID is increased (i.e., an avoidance desire false belief task, see Cassidy, 1998; Leslie, German, & Polizzi, 2005; Leslie & Polizzi, 1998), four-year-olds fail to correctly predict the protagonist’s action. In order to decrease the ID, the saliency of the true belief must be decreased. The saliency of the true belief is reliant on the child’s certainty of the desired object’s whereabouts, and thus, to reduce the ID, the final location of the desired object must be made unknown. This has been implemented in several ways, but only three are described here. First, Bartch (1996) informed the child what the protagonist’s belief was, but then showed the child that both locations were actually empty. Second, Koos et al. (1997) ate the contents of a box in the unexpected contents task before asking the child about the other individual’s false belief. Third, several researchers (Carpenter et al., 2002; Southgate et al., 2007; Wang & Leslie, 2016) removed the desired object “off-screen” to an indefinite location. When ID is decreased in these ways, three-year-olds’ performance improves to “chance,” where “chance” has always been 50% because the responses in the false belief tasks has always been binary (Wellman et al., 2001). Because three-year-olds’ performance only improves to chance, some researchers believe that lowering ID is ineffective and further that
inhibition does not play a role in belief attribution (Devine & Hughes, 2014). In response to Devine and Hughes’ conjecture, Setoh et al. (2016) proposed that there are additional processing demands such as response generation that must be taken into consideration. In addition, although comparing false belief performance to a 50% baseline is the standard in the field, the appropriateness of a 50% comparison is debatable.

A 50% baseline performance is the standard in false belief research. But it is still up for debate whether responses in a false belief task are sampled from a uniform distribution (Carpenter et al., 2002). A 50% baseline is problematic for measuring false belief reasoning because it assumes that the child does not have any pre-existing tendencies. As discussed previously, some argue that children may have the tendency to attribute the true belief (Leslie & Thaiss, 1992; Leslie & Polizzi, 1998), suggesting that a comparison to a 50% baseline does not reflect the true probability distribution for their responses. Indeed, researchers are beginning to deviate from the 50% performance comparison by choosing to compare children’s false belief performance to their performance in a condition that is similar to the false belief condition except the protagonist witnesses the target object switch (Carpenter et al., 2002; Buttelmann et al., 2009; Roth & Leslie, 1998, exp. 1). This condition has been referred to as the “true belief” condition. In the true belief condition, children should predict that the agent would search in the opposite box from the false belief condition. To examine this issue, Carpenter et al. (2002) first compared three-year-olds’ performance on a novel false belief task to the traditional 50% baseline; they found that three-year-olds were “at chance,” but when they compared three-year-olds’ false belief performance to their performance in the true belief condition as a baseline, three-year-olds performed “above
chance.” This suggests that a 50% baseline comparison may not be the most appropriate baseline for examining false belief performance.

1.4. Selecting between performance accounts

In the previous sections, we reviewed two possible performance accounts for why three-year-olds fail the verbal false belief tasks and how correlational studies have supported both accounts. Although some studies have found that both EF components positively correlate to theory of mind, others have not found such correlations. For example, Carlson et al. (2002) found that performance on a battery of WM tasks was not a significant predictor of theory-of-mind performance when age and IQ were controlled for. Indeed, Leslie (1987) suggests that the ability to manipulate two simultaneous representations is present prior to children’s ability to pass the verbal Sally Anne task. This evidence stems from the early emergence of pretend play (Bosco, Friedman, & Leslie, 2006; Onishi, Baillargeon, & Leslie, 2007). When a child engages in pretend play with another individual, the child must decouple the primary/perceptual representation from the metarepresentation. For example, when mom is pretending a banana is a phone, the primary/perceptual representation, Mom is talking to a banana, must be divorced from the metarepresentation, Mom is pretending, “The banana is a phone.” Once divorced, children must flexibly manipulate both representations in order to act appropriately, suggesting that when toddlers engage in pretend play (approximately 15-months), they can already simultaneously keep in mind two representations of the world. Further, another set of findings that is incompatible with the WM performance account is younger children’s success in the spontaneous-response and helping behavior false belief tasks (see p.2). To succeed in these tasks, young children (and infants) must successfully
track both the protagonist’s false belief and their own representation of the desired object. Children’s (and infants’) success in such tasks has been controversial. Although there are many alternative explanations suggesting that children “feign” false belief competence (Apperly & Butterfill, 2009; Heyes, 2014; Perner & Ruffman, 2005; Ruffman & Perner, 2005), there is mounting evidence that rules out these alternative explanations and converge that early false belief reasoning is just as sophisticated as what many deem “real” false belief reasoning (see Scott & Baillargeon, 2017, for a review).

Although the findings from pretense and spontaneous-response false belief tasks are incompatible with the WM performance account, it is still possible that WM demand affects preschoolers’ performance. To present, this possibility cannot be completely ruled out because researchers have never directly tested WM demand in the standard false belief task. Furthermore, it is also unclear whether WM limits preschoolers’ ability to track multiple agents with distinct mental states. Another controversy in EF’s relationship with theory of mind is how EF influences the development of false belief reasoning. Researchers proposed two views on how EF influences theory of mind performance: the emergence view and the expression view. The next section describes each view and its implications.

1.5. Two explanatory accounts for how executive function affects theory of mind development

Like conceptual deficit theorists (Gopnik, 1993; Perner, 1991; Wellman, 1990), “emergentists” argue that young children’s failures on the false belief task are due to a conceptual false belief deficit (Carlson & Moses, 2001; Moses 2001). However, unlike conceptual deficit theorists who argue that “experience” is the catalyst for the acquisition
of the false belief concept, emergentists argue that acquisition of certain EF prerequisites including WM and inhibitory control engenders the concept of false belief. In other words, children require a certain “level” of EF mastery to allow them to entertain perspectives that differ from their own. In support of the emergence view, a wealth of correlational studies have shown executive function scores correlate (Carlson & Moses, 2001; Carlson et al., 2002; Carlson et al., 1998; Frye, Zelazo, & Palfai, 1995; Hughes, 1998a, 1998b) and can even predict later performance on false belief tasks. This holds true even after age (Carlson & Moses, 2001; Hughes & Ensor, 2007) and socioeconomic status (Flynn, O’Malley, & Wood, 2004; Hughes & Ensor, 2005) are accounted for (Carlson, Mandell, & Williams, 2004; Hughes, 1998a; Marcovitch, et al., 2015). One of the major implications of the emergence account is that EF is necessary for the acquisition of theory of mind, but not later in development. If the emergence theory is correct, both adults and older children’s theory of mind performance (i.e., attributing false beliefs) should be impervious to any type of demand.

Unlike the emergence account, the expression account describes EF as a factor that permits the “expression” of theory of mind. According to the expression account, younger children already have the competence to attribute a false belief to another individual, but they fail false belief tasks because their pool of EF resources is insufficient to overcome the inherent EF demands that the traditional false belief task requires from them (Leslie & Polizzi, 1998; Roth & Leslie, 1998). Unlike emergentists, who characterize younger children’s failures as a lack of both WM and inhibitory control mastery, expressionists emphasize the development of inhibitory control for successful belief attribution in false belief tasks. In particular, expressionists argue that in order to
attribute a false belief, subjects must inhibit a salient response, where the salient response in a false belief task is the individual’s own belief (Friedman & Leslie, 2005; Leslie & Polizzi, 1998; Leslie & Thaiss, 1992; Mitchell, 1996). In support of the expression view, adjusting ID alters false belief performance (e.g., increasing ID hinders four-year-olds’ verbal false belief performance, Cassidy, 1998; Friedman & Leslie, 2005; Leslie et al., 2005; Leslie & Polizzi, 1998, and decreasing ID helps three-year-olds’ verbal false belief performance, Bartsch, 1996; Carpenter et al., 2002; Wang & Leslie, 2016). One major implication for the expression view is that the saliency of the true belief and an individual’s ability to inhibit that true belief should continue to affect theory of mind performance throughout development. This should be reflected in individuals committing inhibition (or true belief) errors when EF is taxed.

I have discussed how different EF skills might influence theory of mind, yet differentially testing each account’s predictions is difficult. This is primarily because the two accounts predict the same behavioral response in the standard false belief task – “failure.” The standard false belief task with its single agent only allows a single type of error, precluding error analysis data, a kind of data important in many other areas of cognitive psychology. In order to overcome this limitation, the false belief task must be modified to allow preschoolers the possibility of making different errors. This can be achieved by directly manipulating the WM demand in a false belief task. To manipulate the WM demand in a false belief task, we can increase the number of agents with distinct false beliefs that the preschoolers are required to track.

1.6. Limitations of the current theory of mind tasks
As noted at the outset, three decades of research has examined children’s ability to track only a single agent. Yet, typical social environments, e.g., a small family group, comprise multiple agents and various distinct beliefs and desires. Despite a lack of research in multiple agent theory of mind, researchers sometimes assume that even infants can track multiple agents with distinct mental states, for example, distinct intentions (Choi & Luo, 2015; Hamlin, Wynn, & Bloom, 2007). This assumption sits uneasily with research discussed in Section 1.2 claiming that that WM capacity limits children’s performance in false belief tasks (Gordon & Olson, 1998, Keenan et al., 1998; Olson & Kamawar, 1999).

One advantage of testing children’s theory of mind for multiple agents is that it allows researchers to go beyond the binary “pass/fail” design of the standard task. Whereas there will always be a single way to be fully correct about a given agent’s belief, as the number of agents and beliefs in a task increase, there will be increasingly numerous ways of being wrong. For example, in a double agent false belief task, there are three possible beliefs (two false beliefs and one true belief) and two agents (e.g., Sally and Anne). In such a task, the child can attribute to each agent one of three possible beliefs: Sally’s false belief, Anne’s false belief, and the true belief (the child’s own belief). With two agents and three possible beliefs, subjects can respond from a set of nine possible responses. Having more than two possible responses is useful in two ways. First, it allows more statistical power, for example, in a double agent task the baseline probability is .11 rather than the .5 of the standard single-agent task. The other advantage is that, with eight ways of being wrong, the data can reveal error patterns in a way that
has not been available before. Error analysis is ubiquitous and crucial in other areas of
cognitive psychology, but not in theory of mind so far.

1.7. The current studies: Examining preschoolers’ ability to track multiple agents

By testing preschoolers’ ability to track multiple agents and their distinct false
beliefs, WM demand can be directly studied along with error types. In addition, we can
examine whether “lowering” the ID changes the types of errors that children make. The
following chapters present data on tests of three- and four-year-olds’ ability to track two
agents (Experiments 1 and 2), three agents (Experiments 3 and 4), four agents
(Experiments 5 and 7), and five agents (Experiment 6), each with distinct false beliefs. In
addition, the following chapters present data across two different ID levels, high demand
(Experiment 1, 4, and 7) and low demand (Experiment 2, 3, 5, and 6). These
manipulations are described further in the next chapter.
Chapter 2

Aims of Experiments 1-2

The main aims of the studies in this chapter were 1) to determine whether preschoolers could solve a standard false belief task with two agents each with distinct false beliefs; and 2) to examine which EF skill, WM or inhibition, affected their performance. The first study tested children’s ability to track two agents with distinct false beliefs in a high demand false belief task, where the final location of the desired target was an “on-scene” location. The second study also tested children’s ability to track two agents with distinct false beliefs, but in a low demand false belief task, where the final location of the desired target was an “off-scene” location. If WM influences false belief performance, then preschoolers’ performance should be worse when they are required to track two agents with distinct beliefs than when they are only required to track a single agent. If inhibition influences false belief performance, then preschoolers’ performance should improve when the inhibition demand is low.

Experiment 1: Double Agent, High Demand

To examine whether increasing WM load affects false belief performance, we added an additional agent with a distinct false belief to the standard false belief task. If preschoolers’ performance is limited by their WM capacity, then preschoolers should perform worse on the current task than the standard false belief task. Otherwise, if increasing WM load does not affect preschoolers’ performance, then preschoolers should perform similarly on the current task to the standard false belief task.

Method

Participants
One hundred preschoolers participated in the current study: 50 three-year-olds (29 girls, $M=43$ months, $SD=3$ months) and 50 four-year-olds (23 girls, $M=54$ months, $SD=4$ months). They were recruited from local preschools in Middlesex County, New Jersey. An additional 13 three-year-olds and 8 four-year-olds were tested but were excluded for the following reasons: failing the See question (2 three-year-olds and 2 four-year-olds)$^1$, failing the Reality question (9 three-year-olds and 1 four-year-old), being distracted during the experiment (1 three-year-old), refusing to continue with the story (1 three-year-old), and experimental error, where the experimenter asked the subject a different variant of the test question (4 four-year-olds) and did not continue the story (1 four-year-old).

*Materials*

The story was presented through 10 color pictures in landscape orientation in a three-ring binder. In addition to the color pictures, there were 6 x 2-inch character cutouts for each character. In each image, there were three possible hiding locations: a basket, a box, and a chest. The locations were always in the same position: the basket was located at the lower right-hand corner, the box was located on top of the table at the lower left-hand corner, and the chest was centered at the top of the page. The characters were also always in the same starting position: Sammy (red shirt) on the left and Aaron on the right (see Figure 1).

---

$^1$ Approximately half-way through data collection, children who failed the See question were instead corrected and included in data analysis.
Procedure

Testing occurred in either the child’s preschool classroom or a separate preschool classroom/laboratory room. In all cases, care was taken to ensure that the testing area was quiet. The current experiment was modeled after the original Sally-Anne task with some exceptions: an additional agent with a distinct false belief and an additional location. Like the original Sally-Anne task, the story that children were told was considered a “High Demand” task because the target object’s final location was known by the participant.

Subjects were first introduced to each of the characters, Sammy and Aaron, using the 6-inch character cutouts. Subjects were then shown a picture of Sammy holding the
desired object, a frog, and were asked two warm-up questions: “Do you know what Sammy has in his hand?” and “Do you know what sound a frog makes?” If the subject stated that they are uncertain or they refused to respond, the experimenter asked silly closed ended questions such as, “Is Sammy holding a cat?” or “Does a frog say meow?” Because the child was unfamiliar with the experimenter, we reasoned that these questions would build a more comfortable environment for the subjects. In addition, because of the diversity of the preschools in Middlesex County in New Jersey, an additional purpose for the closed ended questions was to gauge the participant’s English comprehension. They were then told the remainder of the story:

“Sammy was playing with the frog but has to leave, so he puts the frog in the [basket/box]. Then, Sammy leaves. But then Aaron wants to play with the frog, so he takes the frog out of the [basket/box] and plays with it. But now Aaron has to leave, so he puts the frog in the [basket/box]\textsuperscript{2} and then Aaron leaves. But while both the kids were gone, look at what happens! The frog jumps out of the [basket/box] and goes into the chest!”

After the narrative was over, subjects were asked a series of control questions for each agent in order to verify that they understood the story. The first series of control questions examined whether children understood that the agent did not have visual access to the last switch, i.e., “Did Sammy see that?” The second series of control questions was a memory question: “Do you remember where Sammy put the frog?” For each participant, See and Memory questions for both agents were asked. The last control question was a reality control question, “Where is the frog right now?” Finally, subjects were asked to predict where each agent would search for the target object. For each series of questions, questions were ordered either forwards (asking about Sammy first, then Aaron) or reverse (asking about Aaron first, then Sammy).

\textsuperscript{2} The second agent always placed the desired target in a location where the first agent did not.
Coding

Each action prediction response was coded as one of three different belief attributions: Agent 1’s false belief (A1) (that is, the belief that Agent 1 should have, given where Agent 1 put the target object), Agent 2’s false belief (A2) (likewise the belief Agent 2 should have), or the true belief (TB) (that is, the actual final location of the target object). Thus, an A1 response to the prediction question is correct for the first agent, an A2 response is correct for the second agent, and all other responses are incorrect.

Because there were two agents and three possible belief attributions: two false and one true, the number of possibilities is calculated by raising the number of possible beliefs to the power of the number of agents. In the Experiment 1 there are $3^2=9$ possible response patterns, where one pattern is completely correct and the remaining eight patterns are in various ways incorrect.

Results

Figure 2 shows each subject’s individual responses in columns. The top panel (squares) shows four-year-olds and the bottom panel (circles) shows three-year-olds. Twenty-six out of 50 four-year-olds correctly attributed both agents their respective (“fully correct”) false belief (binomial 1/9, $p<.0001$) while 22 out of 50 three-year-olds incorrectly attributed both agents a true belief (binomial 1/9, $p<.0001$). In addition, 10 out of 50 three-year-olds correctly attributed both agents their respective (“fully correct”) false belief (binomial, $p<.05$) and 14 out of 50 four-year-olds attributed a true belief to both agents (binomial 1/9, $p<.001$). No other response patterns occurred significantly often.
Figure 2. Preschoolers’ individual responses for Experiment 1: Double Agent High Demand. Each column of circles indicates a three-year-old participant’s response for Agent 1 (A1) and Agent 2 (A2). Each column of squares indicates a four-year-old’s response for A1 and A2. A fully correct response is attributing Agent 1’s false belief indicated in red to A1 and A2’s false belief indicated in blue to A2.

Question Order: Responses were examined according to question order. A Fisher’s Exact Test revealed that there was no difference in responses across order counterbalances (n.s.). Figure 3 shows subjects’ individual responses grouped by question order.

Figure 3. Preschoolers’ individual responses for Experiment 1: Double Agent High Demand, separated by question order. Each column of circles indicates a three-year-old participant’s response for Agent 1 (A1) and Agent 2 (A2). Each column of squares indicates a four-year-old’s response for A1 and A2. A fully correct response is attributing Agent 1’s false belief indicated in red to A1 and A2’s false belief indicated in blue to A2.

Despite the possibility of nine response patterns, some patterns were not observed. Instead, we identified four different types of responses: All Correct False Belief Attribution, False Belief Binding Error, Mixed False Belief-True Belief Binding Error, and All True Belief Error. All responses were sorted into one of the four categories. To be categorized as an All Correct False Belief Attribution (AC_FB), the
response must include the correct false belief attribution for both of the agents. To be categorized as a False Belief Binding Error (FB_BE), the response must exhibit two characteristics: 1) all attributions were false beliefs and 2) at least one of the false beliefs was attributed to the incorrect agent. For example, a child correctly attributes A1’s false belief to A1 and then incorrectly attributes A1’s false belief to A2. To be categorized as a Mixed False Belief-True Belief Binding Error (MX_BE), the response must contain both false belief and true belief attributions. Finally, to be categorized as an All True Belief Error (A_TBE), the response must contain only true belief attributions. Each of these bins contained a different number of responses. For example, to be categorized in the AC_FB category, children must respond correctly for both agents. This, however, is only one response out of the 9 possible responses. Similarly, to be categorized in the A_TBE category, only one response out of the 9 possible responses fall into this category, i.e., children must attribute to both agents a true belief. For FB_BE and MX_BE, there are three and four responses that could be binned in each category respectively. Figure 4 shows the number of subjects in each of the binned response categories. Although the baseline probabilities for each of the bins varies, we ignored this and conducted a 2 (age) x 4 (response type) chi-square: $\chi^2 (3) = 12.89, p<.005$. The younger preschoolers appeared over all more likely to respond with an A_TBE response than the older preschoolers. Proportions of binned responses are displayed in Figure 4.
Figure 4. Proportion of binned responses for Experiment 1: Double Agent High Demand. Responses were categorized into one of the following four categories: All Correct False Belief Attribution (AC_FB), False Belief Binding Error (FB_BE), Mixed False Belief-True Belief Binding Error (MX_BE), and All True Belief Error (A_TBE). The frequency of each response is displayed above the respective bar.

Finally, to compare preschoolers’ performance in the current task to the traditional single agent false belief task, we further collapsed the responses into two categories. In the traditional task, responses were categorized as one of two possible types: a false belief attribution and a true belief attribution. For false belief attributions, we collapsed the AC_FB and FB_BE responses because both types of responses were all “false belief” attributions. We also included the MX_BE responses that had a false belief for the first belief attribution and a true belief for the second belief attribution because if children were asked to predict only one agent’s action (i.e., the first agent), then they
would have attributed a false belief. This only accounted for two of the 21 MX_BE responses. For true belief attributions, we included the A_TBE responses as well as the subset of MX_BE errors that included a true belief for the first belief attribution and a false belief for the second belief attribution for the same reasoning. Like the standard false belief task, we found that four-year-olds made “false belief” responses (binomial 1/2, 31 out of 50 attributed a “false belief” response, $p<.05$) more often than three-year-olds (binomial 1/2, 14 out of 50 attributed a “false belief” response, $p<.001$) (Yates $\chi^2$ (1)= 10.34, $p=.002$, two-tailed).

Discussion

Four-year-olds perform just as well on a double-agent false-belief task as they do on the standard single-agent task (62% attributing false beliefs), and three-year-olds perform approximately as poorly (only 28% attributing at least one false belief).

Out of nine possible responses, the two most frequent responses for both age groups were attributing the correct false belief to the given agent and attributing the true belief for both agents. Over all, four-year-olds were more likely to attribute correctly each agent his respective false belief while three-year-olds were more likely to attribute incorrectly both agents the true belief. With the baseline probability of 11%, we observed that although three-year-olds as a group were “failing” the verbal false belief task, they were simultaneously “passing” the task as a group by attributing the correct false belief to the given agent. Similarly, we observed that although four-year-olds as a group were “passing” the verbal false belief task, they were also simultaneously “failing” the task as a group by attributing a true belief to both agents. This pattern of results cannot be observed for obvious reasons in the single agent false belief tasks.
In addition to responding with AC_FB and A_TBE, approximately 28% of subjects (18 out of 50 three-year-olds and 10 out of 50 four-year-olds) responded differently. We categorized the responses into meaningful error categories. This enabled the observation of two new types of error, the False Belief Binding Error and Mixed False Belief-True Belief Binding Error. We discuss these further at the end of this chapter.

Despite being required to track an additional agent with a distinct false belief, preschoolers’ performance was strikingly similar to a single agent false belief task. Four-year-olds pass by attributing the correct false belief to each agent and three-year-olds fail, in a systematic way, by attributing a true belief to all of the agents. This finding suggests that WM may not be as influential in theory of mind performance as researchers previously thought. Even in studies with findings that are interpreted as evidence that WM influences theory of mind performance (e.g., Keenan et al., 1998), it is worth noting that according to these studies, three-year-olds’ WM capacity was approximately three items. Further, with a closer examination of the WM and theory of mind studies, even with the largest correlation coefficient ($r=.54$), only 29% of variance is accounted for. If three-year-olds can hold on average at least three items in WM, it seems unlikely that younger children are only capable of tracking a single agent and their belief.

The current study demonstrates that an increase to WM load in the verbal false belief task does not affect preschoolers’ performance much. Although the current study shows that four-year-olds can track two agents without any additional cost to their performance, it is still unknown whether three-year-olds’ poor performance is due to a lack of competence or due to other performance demands such as inhibition.
Additionally, if poor performance is due to inhibition, then can three-year-olds also track both agents and their respective false belief? In Experiment 2, we examined whether three-year-olds can track two agents and their distinct false beliefs in a low ID false belief task.

Experiment 2: Double Agent, Low Demand

There were the same two aims in Experiment 2 as in Experiment 1. First, we wanted to examine whether three-year-olds’ also have the capacity for tracking at least two agents with distinct false beliefs. Second, we wanted to examine if lowering the ID changes three-year-olds’ performance in a double-agent false-belief task. Like previous research (Setoh et al., 2017; Southgate et al., 2007; Wang & Leslie, 2016), we decreased the inhibition demand by removing the target at the end of the story to an indefinite off-scene location.

Method

Participants

One hundred preschoolers participated in the current study: 50 three-year-olds (27 girls, $M=43$ months, $SD=3$ months) and 50 four-year-olds (29 girls, $M=55$ months, $SD=3$ months). They were recruited from local preschools in Middlesex County, New Jersey. An additional 8 three-year-olds and 5 four-year-olds were tested but were excluded for the following reasons: failing the See question (1 three-year-old and 2 four-year-olds).³

³ Half of the participants were originally tested by Dr. Lu Wang, a former graduate student in our lab. This data is reported in an unpublished doctoral dissertation (Wang, 2014). The remaining half of the subjects were tested by the author of this dissertation using the same stimuli and procedure.

⁴ For the participants that Dr. Lu Wang tested, children who failed the See question were not tested further. The remaining participants, who tested by the current author, were instead corrected if they failed the See question and continued with the study.
failing the Reality question (4 three-year-olds and 2 four-year-olds), being a non-native speaker of English (2 three-year-olds), and experimental error, where the experimenter did not ask the see questions (1 three-year-old) and reality question (1 four-year-old).

Materials

The story presented through 13 color pictures in a three-ring binder and in landscape orientation. This was reported in an unpublished doctoral dissertation (see Wang, 2014). We decided to use this story because the former graduate student, Dr. Lu Wang had already tested 20 children and the task was similar to Experiment 1’s approach (i.e., a verbal storybook). There were two possible hiding locations: a red box (always on the left side) and a yellow box (always on the right side). There were two characters, who were always in the same starting position: Sally (in the green shirt) on the left and Anne (in the blue shirt) on the right. There was also a cat, which sat on the mat next to the
table. See Figure 5.

Figure 5. Layout for Experiment 2 (Double Agent Low Demand). Two agents (Sally in green and Anne in blue) stood behind the table and two possible hiding locations: a red box and a yellow box.

Procedure

The procedure was similar to the previous experiment except the end of the story where the desired item was taken by the cat to an indefinite “off-scene” location. See narrative below:

“Look! Here’s Sally and here’s Anne. Do you know what Sally has in her hand? A cookie that’s right! Sally has a cookie, but she has to go away. So she puts the cookie in the [red box/yellow box]. Then, Sally leaves. But then Anne wants the cookie, so she takes the cookie out of the [red box/yellow box]. But now Anne
has to leave, so she puts the cookie in the [red box/yellow box] and then Anne leaves. But while both the kids were gone, who’s left in the room? The cat! That’s right! So the cat smells the delicious cookie. It climbs up onto the table and opens the [red box/yellow box]. Then, it takes the cookie and runs off with it! ”

**Coding**

Like before, there were three possible belief attributions: Agent 1’s false belief (A1), Agent 2’s false belief (A2), and the true belief (TB). Because the current experiment was a “Low Demand” task, the desired target’s location was unknown and the subject’s belief of the desired target’s location was uncertain. So, unlike the previous experiment where indicating the object’s current location was coded as a true belief, responses such as “I don’t know,” “nowhere,” and “everywhere” were coded as true belief attributions. In addition, if the subject generated a location for the desired target (e.g., the cat is outside), the subject’s response was also coded as a true belief attribution. See Chapter 5 for more detail.

Again, we calculated the number of possible responses by raising the number of possible beliefs to the power of number of agents, producing $3^2=9$ possible response patterns, of which only one is completely correct.

**Results**

Figure 6 again shows each subject’s individual responses in columns. The top panel (squares) shows four-year-olds and the bottom panel (circles) shows three-year-olds. Four-year-olds (32 out of 50) correctly attributed both agents their respective (“fully correct”) false belief (binomial 1/9, $p<.001$). When ID was reduced, three-year-olds (20 out of 50) were also able to correctly attribute both agents their respective false beliefs (“fully correct”, binomial 1/9, $p<.001$). While both age groups “passed” by attributing the

---

5 The second agent always placed the desired target in a location where the first agent did not.
correct false belief to the given agent, both age groups also “failed” by attributing both agents a true belief (12 out of 50 four-year-olds, binomial 1/9, $p<.01$ and 21 out of 50 three-year-olds, binomial 1/9, $p<.001$). No other response patterns occurred significantly often.

Figure 6. Preschoolers’ individual responses for Experiment 2: Double Agent Low Demand. Each column of circles indicates a three-year-old participant’s response for Agent 1 (A1) and Agent 2 (A2). Each column of squares indicates a four-year-old’s response for A1 and A2. A fully correct response is attributing Agent 1’s false belief indicated in red to A1 and A2’s false belief indicated in blue to A2.

Question Order: Responses were examined according to counterbalance (forward questions vs. reverse questions). A Fisher’s Exact Test revealed that performance differences in the forwards questioning order and the reverse questioning order were not statistically significant (threes: $p=.08$, fours: $p=n.s.$). See Figure 7.

Figure 7. Preschoolers’ individual responses for Experiment 2: Double Agent Low Demand, separated by question order. Each column of circles indicates a three-year-old participant’s response for Agent 1 (A1) and Agent 2 (A2). Each column of squares indicates a four-year-old’s response for A1 and A2. A fully correct response is attributing Agent 1’s false belief indicated in red to A1 and A2’s false belief indicated in blue to A2.
Preschoolers’ responses were collapsed into the same four categories from Experiment 1. Proportions of binned responses are illustrated in Figure 8. A 2 (age) x 4 (response type) chi-square this time showed no significant difference between the age groups at α=.01: \( \chi^2(3) = 7.02, p=.07, \text{n.s.} \).

![Double False Belief Low Demand: Binned Responses](image)

Figure 8. Proportion of binned responses for Experiment 2: Double Agent Low Demand. Responses were categorized into one of the following four categories: All Correct False Belief Attribution (AC_FB), False Belief Binding Error (FB_BE) Attribution, Mixed False Belief-True Belief Binding Errors (MX_BE), and All True Belief Error (A_TBE). The frequency of each response is displayed above the respective bar.

Like Experiment 1, we then collapsed preschoolers’ responses to binary categories: a “false belief” response or a “true belief” response. Three-year-olds were at chance (binomial 1/2, 25 out of 50 made a “false belief” response, \( p=n.s. \)) while four-year-olds were above chance (binomial 1/2, 37 out of 50 made a “false belief” response,
Chi-square analysis showed this difference to be marginally significant (Yates \( \chi^2 \) (1) = 5.14, \( p = .023 \), two-tailed).

Comparing preschoolers’ performance in the current study to preschoolers’ performance in Experiment 1, a Fisher’s Exact Test revealed that three-year-olds made more all correct responses when ID was lowered than when ID was high (marginally significant, \( p < .02 \)); four-year-olds’ performance did not differ between experiments (n.s.).

Discussion

When the ID was lowered, three-year-olds were also able to track two agents with distinct false beliefs simultaneously. The most powerful analysis (using all the information) showed that with a binomial baseline \( P = .11 \), 40\% of three-year-olds were able not only able to attribute false beliefs but able to do so to the correct agents (\( p < .001 \)). Like Experiment 1, the two most frequent responses for both age groups were attributing both agents their respective false belief and attributing both agents a true belief. Strikingly, we also observed not only that a significant proportion of both age groups were “passing” but at the same time a significant proportion of both age groups were also “failing”!

When preschoolers’ performance was categorized simply as binary responses (throwing away information), four-year-olds, again, were above chance in making a “false belief” response. Three-year-olds, on the other hand, were at chance. Previous research has also found that when demands are “lowered” by removing the desired object to an indefinite location, three-year-olds’ performance only improves to “chance,” where chance probability is assumed to be 50\% (Wellman et al., 2001). Because lowering ID
only results in “chance” performance, researchers have denied that ID is involved in the
standard verbal false belief task (Devine & Hughes, 2014). Devine and Hughes’
reasoning, however, is flawed. It is undeniable that there is a change in performance
when ID is manipulated (below chance performance in the “high demand” tasks to at
chance performance in “low demand” tasks), and yet, the authors do not provide an
account for why the change occurs. It is possible that three-year-olds’ performance
improves to “chance” because lowering ID does not entirely reduce all the inhibition and
the processing demands required to pass the task, which results in only a 20% increase of
three-year-olds “passing” (e.g., see Setoh et al., 2016 on response generation as an
additional processing demand in false belief tasks).

General Discussion Experiments 1-2

The findings in Experiments 1 and 2 demonstrated that tracking an additional
agent did not affect preschoolers’ performance all that much. If a single agent false belief
task could be passed by a child – high ID and low ID for four-year-olds and low ID but
not high ID for three-year-olds – then the child is likely to pass a task with two agents
with distinct false beliefs. Increasing the WM demand makes a small difference in
preschoolers’ performance. In contrast, when the ID was lowered, three-year-olds’
performance changed drastically. These findings suggest that three-year-olds’ poor false
belief performance on the standard verbal single agent false belief task is unlikely due to
the task’s WM demand and is more likely due to the task’s ID.

Not only did Experiments 1 and 2 elucidate the effects of inhibition and WM on
theory of mind performance, Experiments 1 and 2 also provided some clarity on whether
three-year-olds’ failures are caused by a performance problem or a competence problem.
According to the classical conceptual change view, three-year-olds fail the standard verbal single agent false belief task because they lack the concept of false belief (Gopnik, 1993; Perner, 1991; Wellman, 1990). In particular, they suggest that children undergo a qualitative change in their existing belief concept around their fourth birthday. Once children gain the concept of belief, they should demonstrate ceiling performance for all the following time points when their false belief reasoning is tested. On the other hand, the theory of mind mechanism view claims that three-year-olds fail because the task demands hinder three-year-olds’ performance (Leslie & Thaiss, 1992; Roth & Leslie, 1998).

The former view has been heavily favored because preschoolers behave robustly in the standard single agent task: nearly 70% of four-year-olds typically succeed on the verbal false belief task while only 30% of three-year-olds typically succeed on the same task (Wellman et al., 2001). However, because most of the results for the traditional tasks were restricted to a single action-prediction response and limited to a single time point, it was unclear if the standard result, “70% of four-year-olds pass the verbal false belief task,” really indicates a conceptual change type performance. For example, if the result “70% of four-year-olds pass the verbal false belief task” is due to conceptual change, only 7 out of 10 four-year-olds will always pass a false belief task. The result, “70% of four-year-olds pass the verbal false belief task” can also indicate a different type of performance. For example, the result, “70% of four-year-olds pass the verbal false belief task,” could also be generated by a four-year-old who will pass 7 out of 10 times they are administered the false belief task (see Baker, Leslie, Gallistel, & Hood, 2016; Flynn, 2006 for further discussion). Indeed, studies that have examined theory of mind
performance longitudinally have observed that false belief performance is highly variable (Amsterlaw & Wellman, 2006; Flynn, 2006; Flynn, O’Mailey, & Wood, 2004; Mayes, Klin, Tercyak, Cicchetti, & Cohen, 1996) and only a minority of the sample demonstrate performance predicted by a conceptual change account (Baker et al., 2016; Flynn, 2006; Flynn et al., 2004).

Because of the nature of Experiments 1 and 2 where children are required to make two action prediction responses, we can examine how reliable children’s false belief performance is in a single session. The majority of children responded with either an all correct false belief response or an all true belief response. A minority of the sample (28% of preschoolers in Experiment 1 and 15% of preschoolers in Experiment 2) responded differently. Of these preschoolers, 7% of preschoolers in Experiment 1 and 10% of preschoolers in Experiment 2 made a False Belief Binding Error; the remaining preschoolers made a Mixed False Belief-True Belief Binding Error (21% and 5% of preschoolers in Experiments 1 and 2 respectively). Unlike the other responses, the Mixed False Belief-True Belief Binding Error indicates both “failing” and “passing” characteristics, and upon further inspection of the Mixed False Belief-True Belief Binding Errors, the majority of the errors appear systematic. When preschoolers’ Mixed False Belief-True Belief Binding Errors were separated by question order, all but two preschoolers attributed the first agent a true belief and the second agent a false belief. Three-year-olds most noticeably responded as such when ID was high (Experiment 1). With a single test question such as the traditional false belief task, this type of error would be considered “failing” because only the first response (the true belief response) is ever recorded. With a second action-prediction question, we revealed a different type of
response that can inform the cognitive processes underlying theory of mind. Although this type of error appears systematic, it was not statistically significant.

In the next chapter, preschoolers’ WM capacity for tracking multiple agents is explored further. Previous research in WM has suggested that adults can retain only 3-5 information chunks (idioms: Gilchrist, Cowan, & Naveh-Benjamin, 2008; objects: Pylyshyn & Storm, 1988; Scholl & Pylyshyn, 1999; spatial locations: Sanders, 1968). Some researchers have also suggested that adults tracking individuals in conversations may be limited in the same way (3-5 individuals) (Cohen, 1971), which can be observed in media that tries to mimic real social interactions such as Shakespearean plays (Stiller, Nettle, & Dunbar, 2005), soap operas (Matthews & Barrett, 2005), and modern-day films (Krems & Dunbar, 2013). To test multiple agent tracking empirically, we tested preschoolers’ ability to track three agents and their distinct false beliefs across two levels of ID.
Chapter 3

Aims of Experiments 3-4

The main aims for the following two studies were again: 1) to test whether preschoolers can reason about multiple agents each with distinct false beliefs, this time three agents instead of two; and 2) to examine and compare the effect of agent-belief number with ID, testing a low ID task (Experiment 3) and a high ID task (Experiment 4).

Experiment 3: Triple Agent, Low Demand

For Experiment 3 (Triple Agent Low Demand), we increased WM demand by adding an additional agent with a distinct false belief.

Method

Participants

One hundred preschoolers participated in the current study: 50 three-year olds (32 girls, $M=42$ months, $SD=4$ months) and 50 four-year-olds (25 girls, $M=53$ months, $SD=4$ months). They were recruited from local preschools in Middlesex County, New Jersey. An additional 28 three-year-olds and 18 four-year-olds were tested but were excluded for the following reasons: failing the See question (3 three-year-olds and 1 four-year-old)$^6$, failing the Reality question (5 three-year-olds), being a non-native speaker of English (1 three-year-old), being distracted by preschool teachers/classmates or were uncooperative (1 four-year-old), not responding to the questions (1 three-year-olds and 3 four-year-olds), stated that she (the subject) was taking some form of medication (1 three-year-old), and for experimental error, where the experimenter asked the same question more than

---

$^6$ Approximately half-way through data collection, children who failed the See question were instead corrected and included in data analysis.
three times (7 three-year-olds and 4 four-year-olds), asked the subject a different variant of the test question (1 three-year-old and 7 four-year-olds), or did not continue the story (9 three-year-olds and 2 four-year-olds).\(^7\)

**Materials**

The story was presented through 13 color pictures in a three-ring binder and in landscape orientation. In addition to the color pictures, there were 6 x 1-inch character cutouts for each character. The locations (a basket, a box, and a chest) were in the same positions as Experiment 1. There were three characters, who were always in the same starting position: Sammy (red shirt) in the center, Aaron (blue shirt) on the left of Sammy, and Jake (white shirt) on the right of Sammy (see Figure 9).

---

\(^7\) We did not continue with the story for children who failed at least one of the Memory questions. Because the task is a WM task, we realized that we should include these children. Thus, we decided approximately half-way through data collection to continue with the story even after the child failed twice.
Figure 9. Layout for Experiment 3 (Triple False Belief Low Demand). Three agents (Sammy in red, Aaron in blue, and Jake in white) stood in the middle of a triangle formed by three possible hiding locations (a basket, a box, and a chest).

**Procedure**

The procedure was the same as the previous experiment, but with two changes: 1) children were required to track three agents with distinct false beliefs and 2) children were told that the desired object (the frog) “jumped away, far far away, where nobody knows where it went.”

**Coding**

Each action prediction response was coded as one of four different belief attributions: Agent 1’s false belief (A1) (that is, the belief that Agent 1 should have given where Agent 1 put the target object), Agent 2’s false belief (A2) (that is, the belief Agent
2 should have), Agent 3’s false belief (A3) (that is, the belief Agent 3 should have), or the true belief (TB). Because the current experiment is a “Low Demand” task, responses that indicated that the child attributed their own belief were coded as a “true belief” attribution. See the Coding section for Experiment 2.

There were four different possible belief attributions: Agent 1’s false belief (A1), Agent 2’s false belief (A2), Agent 3’s false belief (A3), or the true belief (TB). Again, we calculated the number of possible responses by raising the number of possible beliefs to the power of the number of agents, making $64 (4^3=64$ possible responses). Out of the 64 possible responses, only one response is completely correct.

**Results**

Figure 10 shows each subject’s individual responses in columns. The top panel (squares) shows four-year-olds and the bottom panel (circles) shows three-year-olds. Instead of using the typical $\alpha=.05$, we decided to use a more conservative alpha level because the baseline probability for any response already exceeds 1 in 20, namely, 1 in 64. Analyses for preschoolers’ individual responses were required to meet $\alpha=.0001$. Both three- and four-year-olds in this task correctly attributed all three agents their respective (“fully correct”) false beliefs (binomial 1/64, threes: 20 out of 50, $p<.0001$, and fours: 28 out of 50, $p<.0001$). Three- and four-year-olds also frequently attributed all of the agents a true belief (binomial 1/64, threes: 15 out of 50, $p<.0001$, and fours: 11 out of 50, $p<.0001$).
Figure 10. Preschoolers’ individual responses for Experiment 3: Triple Agent Low Demand. Each column of circles indicates a three-year-old participant’s response for Agent 1 (A1), Agent 2 (A2), and Agent 3 (A3). Each column of squares indicates a four-year-old’s response for A1, A2, and A3. A fully correct response is attributing Agent 1’s false belief indicated in red to A1, A2’s false belief indicated in blue to A2, and A3’s false belief indicated in green to A3.

Question Order: Responses were examined according to counterbalance (forward questions vs. reverse questions). There was no difference in responses across order counterbalances (n.s.). Figure 11 shows subjects’ individual responses grouped by order.

Figure 11. Preschoolers’ individual responses for Experiment 3: Triple Agent Low Demand, separated by question order. Each column of circles indicates a three-year-old participant’s response for Agent 1 (A1), Agent 2 (A2), and Agent 3 (A3). Each column of squares indicates a four-year-old’s response for A1, A2, and A3. A fully correct response is attributing Agent 1’s false belief indicated in red to A1, A2’s false belief indicated in blue to A2, and A3’s false belief indicated in green to A3.

Responses were binned into the same categories as the previous experiments. The proportion of responses for each category is shown in Figure 12. We compared these data
by calculating a 2 (age) x 4 (response type) Chi-square using \( \alpha=0.01; \chi^2 (3) = 5.81, p=0.12, n.s. \). Both age groups responded similarly over all.

![Bar chart](image)

Figure 12. Proportion of binned responses for Experiment 3: Triple Agent Low Demand. Responses were categorized into one of the following four categories: All Correct False Belief Attribution (AC_FB), False Belief Binding Error (FB_BE), Mixed False Belief-True Belief Binding Error (MX_BE), and All True Belief Error (A_TBE). The frequency of each response is displayed above the respective bar.

Because the task is a WM experiment, we realized (eventually!) that we did not want to treat the memory questions as control questions (as is common in the standard Sally and Anne literature), and the results from the children who failed the memory questions were included in all of our analyses. Figure 13 illustrates proportion of responses for both children who successfully answered the memory questions and children who unsuccessfully answered the memory questions. Out of 100 preschoolers,
10 preschoolers answered at least one of the memory questions incorrectly even in the second telling of the story (6 three-year-olds and 4 four-year-olds). In general, there were no systematic patterns found in neither the memory errors committed nor the types of responses made by preschoolers who committed memory errors.

Figure 13. Proportion of binned responses separated by accuracy of memory recall for Experiment 3: Triple Agent Low Demand. Responses were categorized into one of the following four categories: All Correct False Belief Attribution (AC_FB), False Belief Binding Error (FB_BE), Mixed False Belief-True Belief Binding Error (MX_BE), and All True Belief Error (A_TBE).
Preschoolers’ responses were categorized as either a “false belief” or “true belief” attribution. Responses that were categorized as a “false belief” attribution included responses that contained mostly false belief attributions. These responses included AC_FB and FB_BE responses, which only had false belief attributions, and the subset of MX_BE responses that had two out of the three belief attributions as false beliefs.

Responses that were categorized as mostly “true belief” included responses that contained mostly true belief attributions such as A_TBE, and the subset of MX_BE responses where two out of the three belief attributions were true beliefs. Both three- and four-year-olds responded more frequently with a mostly “false belief” response (threes: 33 out of 50, fours: 38 out of 50, binomial 1/2, ps<.01).

Discussion

In Experiment 3, we again find that with a low ID task both three- and four-year-olds successfully track multiple agents each with distinct false beliefs, and can do so for three agents. At the same time, we find again a significant proportion of both age groups attribute a true belief to all agents. Compared to a binary metric (passing by attributing mostly false beliefs or failing by attributing mostly true beliefs), both three- and four-year-olds were considered to be mostly attributing false beliefs. For three-year-olds, this is a surprising finding because previous studies have found that three-year-olds’ performance only improves to chance in low ID tasks (single agent as well as double agent, see Experiment 2). In Experiment 2, when three-year-olds were required to track two agents with distinct beliefs in a low ID task, three-year-olds’ performance compared to a baseline probability of 1/2 was at chance. This could at least in part be due to the increased power of a 1/64 baseline probability. This is supported by the “other half” of
the finding: when described in the usual age group terms, the three-year-olds both “passed” (All Correct category) and “failed” the task (All True Belief category). But exactly the same thing can be said about the four-year-olds, they too both “passed” and “failed” the task. Clearly, in regards to the standard Sally and Anne task with a single agent, such a result is logically impossible!

However, part of this difference (between double and triple agent low ID tasks) could be due to how ID was manipulated in each experiment. In Experiment 2, children were told that the desired target (the cookie) was taken by a cat, whereas, in Experiment 3, children were told that the desired target (the frog) “hopped away, far, far away where nobody knows where it went.” The purpose of removing the desired target to an indefinite location is to reduce child’s certainty of the desired target’s final location. Both manipulations ultimately reduced the saliency of the child’s own belief of the desired target’s final location, but there is a slight difference in the certainty about the desired target’s final location. For Experiment 2, when asked to predict where the agent will search for the cookie, there is some certainty of the cookie’s final location because preschoolers know that the cat has the cookie, but they are not certain about the cat’s final location. In contrast, for Experiment 3, there is a greater uncertainty about the frog’s location because the frog’s final location is completely unspecified. Therefore, the final location of the desired object was vaguer in Experiment 3 than Experiment 2, which may have resulted in an “easier” or a “lower ID” task.

Experiment 4: Triple Agent, High Demand
In Experiment 4, we changed the desired toy from a frog to a chick because preschoolers in Experiments 1 and Experiment 3 frequently confused the frog for an airplane, which might be due to its geometric shape. It is possible that preschoolers may have believed that the airplane was being operated off-scene by one of the characters and may have affected the results. Because we did not find any significant differences in counterbalancing the question order for Experiments 1-3, we only asked the questions in the forward order.

Method

Participants

One hundred preschoolers participated in the current study: 50 three-year olds (32 girls, \(M=42\) months, \(SD=3\) months) and 50 four-year-olds (25 girls, \(M=54\) months, \(SD=3\) months). They were recruited from local preschools in Middlesex County, New Jersey. An additional 26 three-year-olds and 15 four-year-olds were tested but were excluded for the following reasons: failing the See question, (2 four-year-olds), failing the Reality question (14 three-year-olds and 7 four-year-olds), being a non-native speaker of English (2 three-year-old), being distracted by preschool teachers/classmates or were uncooperative (6 three-year-olds), refusing to listen to the story a second time (1 three-year-old), not responding to the questions (3 three-year-olds and 3 four-year-olds), and for experimental error, where the experimenter ran the incorrect counterbalance (1 four-year-old).

Materials

---

8 Initially, about half of the participants were screened for not passing the See question, and therefore did not continue the task if they failed the See question. The remaining half of the participants were corrected for the See question.
The story was presented through 13 color pictures in landscape orientation in a three-ring binder. In addition to the color pictures, there were 6 x 1-inch character cutouts for each character. The possible hiding locations for the desired object included the previous three locations (the basket, the box, and the chest) and a new location, a jar. Like the previous experiments, the locations were always in the same positions and fixed. But this time in a diamond shape: the basket was located on the center right of the page, the jar was located on the center of the bottom of the page, the box was on top of the table on the center left of the page, and the chest was centered on the top of the page. The three characters (Sammy, Aaron, and Jake) were in the same position as the previous experiment (see Figure 14).
Figure 14. Layout for Experiment 4 (Triple Agent High Demand). Three agents (Sammy in red, Aaron in blue, and Jake in white) stood in the middle of a diamond formed by four possible hiding locations (a basket, a jar, a box, and a chest).

**Procedure**

Subjects were tested using the same procedure as Experiment 3 except that preschoolers were told that the chick flew into the last remaining location. Below is the story narrative:

“Sammy was playing with the chick but has to leave, so he puts the chick in the [basket/box/jar/chest]. Then, Sammy leaves. But then Aaron wants to play with the chick, so he takes the chick out of the [basket/box/jar/chest] and plays with it. But now Aaron has to leave, so he puts the chick in the [basket/box/jar/chest] and then Aaron leaves. But then Jake wants to play with the chick, so he takes the chick out of the [basket/box/jar/chest] and plays with it. But now Jake has to leave, so he puts the chick in the [basket/box/jar/chest] and then Jake leaves. But while the kids were gone, look at what happens! The chick jumps out of the [basket/box/jar/chest] and goes into the [basket/box/jar/chest]!”

**Coding**

Each action prediction response was coded as before except because the task was a high ID task, a response where a preschooler indicated the actual final location of the target object was coded as the true belief (TB).

The number of possible responses is the same as before: $4^3=64$ possible responses), where only one response is completely correct and where the remaining 63 responses are incorrect in some way or other.

**Results**

Figure 15 shows each subject’s individual responses in columns. The top panel (squares) shows four-year-olds and the bottom panel (circles) shows three-year-olds. We again used $\alpha=.0001$ to examine the significance of the individual responses. A subset of four-year-olds correctly attributed all agents their respective (“fully correct”) false beliefs (17 out of 50, binomial 1/64, $p<.0001$). While there was a subset of four-year-olds
“passing” the task “as a group”, there was a subset of four-year-olds who were simultaneously “failing” the task by attributing a true belief to all of the agents (12 out of 50, binomial 1/64, $p<.0001$). By contrast, only 5 out of 50 three-year-olds were “all correct” (n.s.) in the high ID version. Nineteen out of 50 three-year-olds attributed all agents a true belief (binomial 1/64, $p<.0001$). One other error pattern occurred significantly often: three-year-olds attributed a true belief for the first agent and the correct false belief for the subsequent two agents (7 out of 50 three-year-olds, binomial 1/64, $p<.0001$). No other response patterns occurred at an above chance rate.

Figure 15. Preschoolers’ individual responses for Experiment 4: Triple Agent High Demand. Each column of circles indicates a three-year-old participant’s response for Agent 1 (A1), Agent 2 (A2), and Agent 3 (A3). Each column of squares indicates a four-year-old’s response for A1, A2, and A3. A fully correct response is attributing Agent 1’s false belief indicated in red to A1, A2’s false belief indicated in blue to A2, and A3’s false belief indicated in green to A3.

Responses were binned into the same categories as previous experiments as illustrated in Figure 16. In examining the MX_BE error, one systematic error accounted for the majority of the responses. The majority of preschoolers who committed a MX_BE error attributed the first agent a true belief and the remaining agents a false belief (threes: 19 out of 24 MX_BE and fours: 15 out of 17 MX_BE). We again compared these data by calculating a 2 (age) x 4 (response type) Chi-square using $\alpha=.01$: $\chi^2 (3) = 9.9879$, marginally significant at $p=.0187$). There was a tendency in the high ID version of the
triple agent task, for demand level to differentiate the two age groups, with the 4 year olds passing and the 3 year olds failing.

Figure 16. Proportion of binned responses for Experiment 4: Triple Agent High Demand. Responses were categorized into one of the following four categories: All Correct False Belief Attribution (AC_FB), False Belief Binding Error (FB_BE), Mixed False Belief-True Belief Binding Error (MX_BE), and All True Belief Error (A_TBE). The frequency of each response is displayed above the respective bar.

Again, because the task was testing WM, we included preschoolers who incorrectly recalled any of the agents’ placement of the desired object (i.e., failing the memory question). Out of 100 preschoolers, 25 preschoolers answered at least one of the memory questions incorrectly even in the second telling of the story (17 three-year-olds and 8 four-year-olds). Eight out of 17 three-year-olds who incorrectly recalled at least
one of the agents’ placement of the desired object attributed the first agent a true belief and the remaining agents a false belief. Similarly, five out of 8 four-year-olds who incorrectly recalled at least one of the agents’ placement of the desired object also attributed the first agent a true belief and the remaining agents a false belief. By adding preschoolers who failed the memory recall, it appears that we inflated the proportion of error for the three-year-olds. Results, however, remain the same. See Figure 17 for the proportions of response type for both children who correctly responded to all of the memory questions and children who incorrectly responded to at least one of the memory questions.
Preschoolers’ responses categorized as a false belief response (mostly false belief attributions) or a true belief response (mostly true belief attributions). Three-year-olds were at chance (29 out of 50 made “mostly false belief” responses, n.s.); four-year-olds were above chance (38 out of 50 made “mostly false belief” responses, binomial 1/2,
Chi-square analysis showed this difference was not statistically significant
(Yates $\chi^2(1) = 2.99$, $p=.09$, two-tailed).

Discussion

Experiment 4 showed that in a high-demand triple-agent false-belief task, the two age groups behaved in different ways. Three-year-olds tended to fail by making All True Belief Errors, while four-year-olds tended to attribute the correct false belief to a given agent. Three- and four-year-olds also made more Mixed False Belief-True Belief Errors (41 out of 100 preschoolers) than the previous experiments. In Experiments 1 and 2, we observed that the majority of Mixed False Belief-True Belief Errors appeared to be systematic and that three-year-olds are especially prone to making this error when ID is high. Likewise, we found the same pattern in the Triple Agent High Demand task, where preschoolers tend to attribute the first agent a true belief and then the remaining agents a false belief (though not necessarily the correct false belief). We will refer to this error as the “bleeding true belief error” because inspection of the individual responses figures (for example, Figure 15) gives an impression of the “yellow” responses sliding or bleeding to the left into the other false belief responses.

There are at least two possible explanations for why bleeding true belief errors occur. One possible explanation is that preschoolers are second-guessing their responses. It is possible that asking the child the same question again, albeit for a different agent, causes the child to re-evaluate their initial response and then change for their second response (Moston, 1987; Poole & White, 1991; Warren, Hulse-Trotter, & Tubbs, 1991). According to Krähenbühl, Blades, & Eiser (2009), children who are uncertain about the “correct” answer are especially susceptible to this type of behavior. This explanation can
be ruled out, however. If preschoolers were second-guessing their answers, then we should have observed preschoolers performing randomly. The bleeding true belief errors (and the other prominent error, the All True Belief Error, found in these tasks) are systematic, which is incompatible with the second-guessing rule.

The other account is an inhibitory one. Preschoolers making the bleeding true belief errors because they are on the cusp of having sufficient inhibitory resources. Having additional action-prediction questions may make the prepotent response of attributing the true belief dissipate. For example, when the first action-prediction question is asked, children may be unable to inhibit their initial response to attribute the true belief to the agent. But with an additional action-prediction question, it is possible that the passage of time between the first question to the next action-prediction question affords children “time” to inhibit the prepotent response of attributing the true belief. Time delays in inhibition tasks such as the Day/Night task have previously been found to be beneficial for four-year-olds (Diamond, Kirkham, & Amso, 2002; Gerstadt et al., 1994; Ling, Wong, & Diamond, 2016). In the Day/Night task, four-year-olds’ performance is mediocre, but when the experimenter chants a short song before the child is able to respond, four-year-olds’ accuracy increases dramatically (Diamond et al., 2002; Gerstadt et al., 1994; Ling et al., 2016).

Both explanations predict that children respond systematically where the first response is a true belief and the remaining a false belief. When we separate children’s Mixed False Belief-True Belief Errors by question order for Experiments 1-3 and examine children’s Mixed False Belief-True Belief Errors in Experiment 4, the majority
of these errors are systematic in this way. Because both explanations predict the same systematic response, we are unable to favor one explanation over the other.

General Discussion for Experiments 3-4

Experiments 3 and 4 have shown that four-year-olds can keep track of at least three agents with distinct false beliefs and three-year-olds can also keep track of at least three agents with distinct false beliefs as long as the ID is lowered. This finding aligns with previous research showing that adults’ general WM capacity (for conversations: Cohen, 1971, for objects: Pylyshyn & Storm, 1988; Scholl & Pylyshyn, 1999; for spatial locations: Sanders, 1968) and adults’ specific WM capacity for tracking individuals (Cohen, 1971) are limited to approximately 3-5 chunks of information.

The upper limit of WM capacity for tracking multiple agents with distinct false beliefs is probed in Chapter 4 with four-agents (Low Demand: Experiment 5, High Demand: Experiment 7) and five agents (Experiment 6). In addition, we wanted to observe the types of errors preschoolers make if both WM and inhibition were taxed. Although Experiments 1-4 have answered which EF plays an important role for false belief reasoning, we have yet to answer how EF affects false belief reasoning. Both the “emergence” and the “expression” accounts in theory of mind development agree when preschoolers should succeed and when they should fail the standard false belief task, they disagree on why preschoolers perform the way they do and on the types of errors that children would make if EF is taxed. For the emergence account, since EF is only necessary for the acquisition of theory of mind, emergentists predict that even if EF is taxed, preschoolers should perform similarly to the standard single agent false belief task. That is, four-year-olds should pass by attributing only false beliefs and three-year-olds
should fail by attributing true beliefs. In other words, three-year-olds should continue to fail by attributing true beliefs to the agents while four-year-olds should continue to attribute false beliefs. In contrast, for the expression account, since EF is described as a pool of resources used in the verbal false belief task, expressionists predict that both three- and four-year-olds should make true belief errors when EF is taxed.
Chapter 4

4.0 Aims for Experiments 5-7

There were two aims for Experiments 5-7. The first was to probe the upper limit of preschoolers’ WM capacity for tracking multiple agents with distinct false beliefs. The second was to examine how both high WM demand in tandem with high ID affects preschoolers’ false belief performance.

Experiment 5: Quadruple Agent, Low Demand

The aim of the current study was to examine whether theory of mind WM capacity is similarly limited as previous research in individual tracking (Cohen, 1971; Matthews & Barrett, 2005; Krems & Dunbar, 2013; Stiller et al., 2005) and research in other domains (Gilchrist et al., 2008; Pylyshyn & Storm, 1988; Sanders, 1968; Scholl & Pylyshyn, 1999). Will we see a catastrophic decline in performance with 4 or 5 agents or a gentle degradation?

Method

Participants

One hundred preschoolers participated in the current study: 50 three-year olds (28 girls, $M=42$ months, $SD=3$ months) and 50 four-year-olds (29 girls, $M=56$ months, $SD=3$ months). They were recruited from local preschools in Middlesex County, New Jersey. An additional 15 three-year-olds and 4 four-year-olds were tested but were excluded for the following reasons: failing the Reality question (5 three-year-olds), being a non-native speaker of English (1 three-year-old), being distracted by preschool teachers/classmates (1 three-year-old), refusing to listen to the story a second time (1 three-year-old and 1 four-year-old), not responding to any of the experimenter’s questions (6 three-year-olds),
and for experimental error, where the experimenter mixed two of the characters’ names (3 four-year-olds).

Materials

The story was presented through 16 color pictures in landscape orientation in a three-ring binder. In addition to the color pictures, there were 6 x 1-inch character cutouts for each character. The locations were the same (a basket, a jar, a box, and a chest) and were in the same positions as Experiment 4. The characters included two boys and two girls; all of the characters were created to be perceptually distinct from one another. The starting position of the characters formed a diamond where Sammy (in red) was on the bottom corner, Aaron (in blue) was on the left corner, Jane (in green) was on the right corner, and Mary (in pink) was on the top corner (see Figure 18).
Figure 18. Layout for Experiment 5 (Quadruple Agent Low Demand). Four agents (Sammy in red, Aaron in blue, Jane in green, and Mary in pink) stood in the middle of a diamond formed by possible hiding four locations (a basket, a jar, a box, and a chest).

Procedure

The procedure was the same as the previous experiment except for two changes: 1) children were required to track four agents with distinct false beliefs and 2) children were told that the chick “flew away, far far away, where nobody knows where it went.”

Coding

Each action prediction response was coded as before but with an additional agent, agent 4 (A4). With five different possible belief attributions (four false and one true), the number of possible responses a subject could make 625 was \(5^4 = 625\) possible responses. Out of the 625 possible responses, only one response was completely correct.

Results

Figure 19 shows each subject’s individual responses in columns. The top panel (squares) shows four-year-olds and the bottom panel (circles) shows three-year-olds. Like the previous experiments, because the baseline probability=.0016, we used \(\alpha=.000001\) to evaluate “significantly” different responses. Both three- and four-year-olds continued to successfully attribute all agents their respective (“fully correct”) false belief (binomial 1/625, threes: 11 out of 50, \(p<.000001\) and fours: 25 out of 50 four-year-olds, \(p<.000001\)). Like the previous studies, three- and four-year-olds also frequently attributed true beliefs to all of the agents (binomial 1/625, threes: 19 out of 50, \(p<.000001\) and fours: 16 out of 50, \(p<.000001\)). No other response patterns occurred significantly often at \(\alpha=.10^6\).
Figure 19. Preschoolers’ individual responses for Experiment 5: Quadruple Agent Low Demand. Each column of circles indicates a three-year-old participant’s response for Agent 1 (A1), Agent 2 (A2), Agent 3 (A3), and Agent 4 (A4). Each column of squares indicates a four-year-old’s response for A1, A2, A3, and A4. A fully correct response is attributing A1’s false belief indicated in red to A1, A2’s false belief indicated in blue to A2, A3’s false belief indicated in green to A3, and A4’s false belief indicated in pink to A4.

Responses were binned into the same categories as the previous experiments. See for proportion of binned categories in Figure 20. We compared these data by calculating a 2 (age) x 4 (response type) Chi-square using \( \alpha = .01 \): \( \chi^2 (3) = 9.89, p = .019 \), marginally significant). Both age groups responded similarly over all but there was a tendency for four-year-olds to respond more “accurately” than the three-year-olds on AC_FB vs. FB_BE. The other two categories appear to be very similar. Focusing on the former two categories, we conducted a Yates corrected 2 x 2 \( \chi^2 \) analysis: \( \chi^2 (1) = 7.1, p = .008 \), two-tailed. For the first time (quadruple agent), the older preschoolers are outperforming the younger preschoolers on a low-demand task.
Figure 20. Proportion of binned responses for Experiment 5: Quadruple Agent Low Demand. Responses were categorized into one of the following four categories: All Correct False Belief Attribution (AC_FB), False Belief Binding Error (FB_BE), Mixed False Belief-True Belief Binding Error (MX_BE), and All True Belief Error (A_TBE). The frequency of each response is displayed above the respective bar.

Out of 100 preschoolers, 21 preschoolers answered at least one of the memory questions incorrectly even in the second telling of the story (16 three-year-olds and 5 four-year-olds). Ten out of 16 three-year-olds who incorrectly recalled at least one of the agents’ placement of the desired object made an incorrect false belief response. See Figure 21. Inspection of Fig 21, especially the FB Binding error 3-year-olds, shows, for the first time in our studies, a marked uptick in Incorrect Memory Responses – in fact more than half the subjects in this bin have incorrect memory responses. It is likely that
the younger children are beginning to struggle with the WM demands of the task, even though the ID is low. Four-year-olds in contrast appear to be unfazed.

![Bar chart](image)

**Figure 21.** Proportion of binned responses separated by accuracy of memory recall for Experiment 5: Quadruple False Belief Low Demand. Responses were categorized into one of the following four categories: All Correct False Belief Attribution (AC_FB), False Belief Binding Error (FB_BE), Mixed False Belief-True Belief Binding Error (MX_BE), and All True Belief Error (A_TBE).

Again, preschoolers’ responses were categorized as a “mostly” false belief response or a “mostly” true belief response. Previously, a “mostly” false belief response included mostly false belief attributions and a “mostly” true belief response included
mostly true belief attributions. Because there were four belief attributions for each response, it is possible that there is not a majority (i.e., two false belief attributions and two true belief attributions). This accounted for one out of 100 responses (1 three-year-old). This response was excluded from this analysis. Using this binary metric, three-year-olds were at chance (binomial 1/2, 29 out of 49 (59%) made “mostly” false belief responses, n.s.) while four-year-olds were above chance (binomial, 1/2, 34 out of 50 (68%) made “mostly” false belief responses, p<.02, two-tailed).

Discussion

Three-year-olds’ WM capacity for tracking multiple agents may be limited to approximately three to four agents. The proportions of three-year-olds’ “fully correct” false belief responses dropped considerably (2-Agents: 40%, 3-Agents: 40%, 4-Agents: 22%). Unlike previous low demand false belief tasks, three-year-olds in the Four Agent Low Demand task most frequently attributed a true belief to each of the agents. It is worth noting that when responses were categorized as mostly attributing false beliefs or mostly attributing true beliefs, three-year-olds were considered to be “at chance,” which differs from their performance in high demand tasks where three-year-olds are typically “below chance.” This could be due to three-year-olds making a high proportion of False Belief Binding Errors, suggesting that the task’s WM demand may for the first time be impeding three-year-olds’ performance. Four-year-olds, on the other hand, continue to succeed by attributing the correct false belief to the given agent. The proportion of four-year-olds’ “fully correct” false belief responses continue to be stable, but decreasing steadily (2-Agents: 64%, 3-Agents: 56%, 4-Agents: 50%).
In Experiment 6 (Pentuple Agent Low Demand), we investigated four-year-olds’ theory of mind working memory capacity limit and what types of errors they tend to commit if they are overwhelmed by WM demand (assuming that 5 agents will finally sink them).

Experiment 6: Pentuple Agent, Low Demand

The main goal of Experiment 6 is to determine four-year-olds’ theory of mind WM capacity and to examine the types of errors four-year-olds make if their WM capacity limit has been reached. Because previous research has found that adults’ WM capacity for tracking individuals may be approximately four individuals (Cohen, 1971; Matthews & Barrett, 2005; Krems & Dunbar, 2013; Stiller et al., 2005), we suspect that four-year-olds’ WM capacity for tracking multiple agents should also be approximately four to five agents. Because the previous experiment showed that three-year-olds’ WM capacity is limited to approximately four agents, we examined only four-year-olds’ ability to track five agents with distinct false beliefs in a low demand task.

Method

Participants

Fifty four-year-olds (22 girls, \(M=55\) months, \(SD=4\) months) participated in the current study. They were recruited from local preschools in Middlesex County, New Jersey. An additional 10 four-year-olds were tested but were excluded for the following reasons: failing the Reality question (4), being a non-native speaker of English (1), being distracted by preschool teachers/classmates (2), refusing to listen to the story a second
time (2), and for experimental error, where the experimenter switched two of the characters’ names (1).

Materials

The story was presented through 18 color pictures in landscape orientation in a three-ring binder. In addition to the color pictures, there were 6 x 1-inch character cutouts for each character. There were five locations that formed a circle: a box on a table on the left center of the page, a chest at the top center, a cabinet on the right center, a basket at the bottom right, and a jar at the bottom left. In addition to the previous four characters in Experiment 5, the current experiment included another boy, Billy (in yellow). The characters also formed a pentagon (see Figure 22).
Figure 22. Layout for Experiment 6 (Pentuple Agent Low Demand). Five agents (Sammy in red, Aaron in blue, Jane in green, Mary in pink, and Billy in yellow) stood in the middle of a circle formed by five possible hiding locations (a basket, a jar, a box, a chest, and a cabinet).

**Procedure**

The procedure was the same as the previous experiment except with an additional agent.

**Coding**

Each action prediction response was coded as before, but with an additional agent, Agent 5 (A5).

With five possible belief attributions: five false and one true and four agents, there are $6^5 = 7,776$ possible responses a preschooler could make. Out of the 7,776 possible responses, only one response is completely correct.

**Results**

Figure 23 shows each subject’s individual responses in columns. In Experiment 6, 18 out of 50 four-year-olds attributed all true beliefs to all of the agents in the task (binomial 1/7,776, $p < .000001$). Because of the sensitivity of the test, four-year-olds also “significantly” attributed both agents their respective (“fully correct”) false belief (binomial 1/7,776, 9 out of 50 four-year-olds, $p < .00001$) and attributed the last agent’s false belief to all of the agents (binomial 1/7,776, 4 out of 49 four-year-olds, $p < .00001$).
Figure 23. Preschoolers’ individual responses for Experiment 6: Pentuple Agent Low Demand. Each column of circles indicates a three-year-old participant’s response for Agent 1 (A1), Agent 2 (A2), Agent 3 (A3), Agent 4 (A4), and Agent 5 (A5). Each column of squares indicates a four-year-old’s response for A1, A2, A3, A4, and A5. A fully correct response is attributing A1’s false belief indicated in red to A1, A2’s false belief indicated in blue to A2, A3’s false belief indicated in green to A3, A4’s false belief indicated in pink to A4, and A5’s false belief indicated by purple to A5.

Responses were binned into the same categories as the previous experiments. See Figure 24.
Figure 24. Proportion of binned responses for Experiment 6: Pentuple Agent Low Demand. Responses were categorized into one of the following four categories: All Correct False Belief Attribution (AC_FB), False Belief Binding Error (FB_BE), Mixed False Belief-True Belief Binding Error (MX_BE), and All True Belief Error (A_TBE). The frequency of each response is displayed above the respective bar.

Out of 50 preschoolers, 15 four-year-olds answered at least one of the memory questions incorrectly even in the second telling of the story. Seven out of 15 four-year-olds who incorrectly recalled at least one of the agents’ placement of the desired object made an all true belief response. Six out of 15 four-year-olds made a False Belief Binding Error. See Figure 25. With five agents, the False Belief Binding Error category is beginning to contain a sizable proportion of Incorrect Memory Responses, echoing what we observed in the previous experiment with three-year-olds. Also notable is the
proportion in memory errors in the All True Belief bin as well. This suggests that 5-agent stories are beginning to stretch 4-year-olds thin.

Figure 25. Proportion of binned responses separated by accuracy of memory recall for Experiment 6: Pentuple False Belief Low Demand. Responses were categorized into one of the following four categories: All Correct False Belief Attribution (AC_FB), False Belief Binding Error (FB_BE), Mixed False Belief-True Belief Binding Error (MX_BE), and All True Belief Error (A_TBE).

Again, we categorized children’s responses as either a “mostly false belief” response or a “mostly true belief” response. Using this categorization, four-year-olds were attributing mostly “false beliefs” (binomial 1/2, 31 out of 50, $p < .05$)
Discussion

Like three-year-olds with four agents with distinct false beliefs, four-year-olds demonstrated a significant drop-off in “fully correct” false belief performance when they were required to track five agents with distinct false beliefs (2-Agents: 64%, 3-Agents: 56%, 4-Agents: 50%, and 5-Agents: 18%). When four-year-olds were required to track five agents with distinct false beliefs, the most common response was attributing a true belief to all of the agents. When we categorized four-year-olds’ responses as “mostly” false belief responses or “mostly” true belief responses, we found that four-year-olds were attributing mostly “false beliefs.” Like three-year-olds, we observed an increase of False Belief Binding Errors and within these an uptick in Incorrect Memory Responses. Taken together, these findings strongly suggest that four-year-olds’ performance shares similar characteristics to three-year-olds’ in Experiment 5 (Quadruple Agent Low Demand) when WM capacity’s limit was being reached.

With respect to the emergence and expression debate, the current findings are consistent with both views. According to the emergence view, once the concept of false belief has emerged, EF is no longer a factor in false belief performance (Carlson & Moses, 2001; Moses, 2001), suggesting that four-year-olds (whose false belief concept has already emerged) should continue to attribute false beliefs. The current study showed that four-year-olds did attribute false beliefs, albeit the incorrect ones, when the task’s WM demand is high, but ID is low. Experiment 6’s findings are also consistent with the expression view, which argues that EF continues to be a major factor in false belief performance throughout the child’s development, suggesting that three- and four-year-olds’ performance should be similar if EF is taxed enough. To discern which view is
correct, four-year-olds’ false belief performance must be examined in a false belief task where their EF is heavily taxed.

In Experiment 4, we observed that four-year-olds’ performance began to falter when ID was high and they were required to track three agents with distinct false beliefs. Namely, their “fully correct” proportion in the Three Agent High Demand task (34%) was a fairly large drop-off in comparison to their “fully correct” proportion in the Three Agent Low Demand task (56%), suggesting that four-year-olds like three-year-olds appear to be susceptible to ID. we may be able to examine the effects of EF on four-year-olds’ performance if they engage in a task where both ID and WM demand are high. For Experiment 7, we examined four-year-olds’ performance in a Four Agent High Demand task. The emergence view predicts that four-year-olds should continue to attribute false beliefs because their EF mastery has already engendered the concept of false belief. In contrast, the expression view predicts that if four-year-olds’ EF is taxed, four-year-olds’ performance should converge on the same poor performance as three-year-olds.

Experiment 7: Quadruple Agent, High Demand

The main objective for Experiment 7 is to examine whether four-year-olds’ performance would be similar to three-year-olds’ if both WM and ID were high.

The emergence view predicts that four-year-olds should still perform well as defined as attributing “false beliefs.” In contrast, the expression view predicts that four-year-olds, if EF is taxed heavily, should perform poorly by attributing a true belief to the agents in question.

Method
Participants

Fifty four-year-olds (26 girls, $M=54$ months, $SD=3$ months) participated in the current study. They were recruited from local preschools in Middlesex County, New Jersey. An additional 8 four-year-olds were tested but were excluded for the following reasons: failing the Reality question (4), being a non-native speaker of English (1), being distracted by preschool teachers/classmates (2), and for experimental error, where the experimenter mixed two of the characters’ names (1).

Materials

The story was presented through 16 color pictures in landscape orientation in a three-ring binder. In addition to the color pictures, there were 6 x 1-inch character cutouts for each character. The locations were the same as Experiment 6 (basket, box, chest, jar, and cabinet) and the characters used in the current task were the same as Experiment 5 (Sammy, Aaron, Jane, and Mary) (see Figure 26).
Figure 26. Layout for Experiment 7 (Quadruple Agent High Demand). Four agents (Sammy in red, Aaron in blue, Jane in green, and Mary in pink) stood in the middle of a circle formed by five possible hiding locations (a basket, a jar, a box, a chest, and a cabinet).

Procedure

The procedure was the same as Experiment 5 except children were told at the end of the story that the chick flew into the remaining location.

Coding

We coded the responses like in Experiment 5 except true belief responses were always the chick’s current location. Again, with five different possible beliefs (four false and one true) and four agents, there were 625 possible responses that could be made.

Results
Figure 27 shows each subject’s individual responses in columns. Four-year-olds performed strikingly similar to three-year-olds in other high ID tasks: they predominantly responded all true belief for all agents (binomial 1/625, 22 out of 50, $p<.000001$). Again because of the sensitivity of the test, two other individual responses that were significantly above chance: attributing the correct false belief to the given agent (“fully correct,” binomial 1/625, 8 out of 50, $p<.000001$) and attributing a true belief to the first agent and then the correct false belief for the remaining agents (6 out of 50, $p<.00001$).

No other responses occurred significantly often.

Figure 27. Preschoolers’ individual responses for Experiment 7: Quadruple Agent High Demand. Each column of circles indicates a three-year-old participant’s response for Agent 1 (A1), Agent 2 (A2), Agent 3 (A3), and Agent 4 (A4). Each column of squares indicates a four-year-old’s response for A1, A2, A3, and A4. A fully correct response is attributing A1’s false belief indicated in red to A1, A2’s false belief indicated in blue to A2, A3’s false belief indicated in green to A3, and A4’s false belief indicated in pink to A4.

Individual responses were binned into the same categories as previous experiments. Figure 28 shows the proportions for each response category. We can compare four-year-olds on Triple Agent High Demand (see Chapter 3) with Quadruple Agent High Demand (current experiment). Such a comparison isolates the effect of an additional agent while holding demand level constant. Because FB_BE and MX_BE responses were almost numerically identical, we conducted a Yates corrected $2 \times 2 \chi^2$
analysis on the other two types of responses (AC_FB and A_TBE): $\chi^2 (1)=4.93, p = .026$, two-tailed. Four-year-olds made fewer AC_FB responses in the Quadruple Agent High Demand (8 out of 50) task than the Triple Agent High demand task (17 out of 50). In addition, four-year-olds made more A_TBE errors in the Quadruple Agent High Demand task (22 out of 50) than in the Triple Agent High Demand task (12 out of 50).

![Figure 28. Proportion of binned responses for Experiment 7: Quadruple Agent High Demand. Responses were categorized into one of the following four categories: All Correct False Belief Attribution (AC_FB), False Belief Binding Error (FB_BE), Mixed False Belief-True Belief Binding Error (MX_BE), and All True Belief Error (A_TBE). The frequency of each response is displayed above the respective bar. Out of 50 preschoolers, 9 four-year-olds answered at least one of the memory questions incorrectly even in the second telling of the story. Three out of 9 four-year-olds](image-url)
who incorrectly recalled one of the agents’ placement of the desired object made an all true belief response. See Figure 29.

Figure 29. Proportion of binned responses separated by accuracy of memory recall for Experiment 7: Quadruple Agent High Demand. Responses were categorized into one of the following four categories: All Correct False Belief Attribution (AC_FB), False Belief Binding Error (FB_BE), Mixed False Belief-True Belief Binding Error (MX_BE), and All True Belief Error (A_TBE).

We categorized responses as a binary response: “mostly” false belief response or “mostly” true belief response. Two responses did not contain a majority, and therefore
were excluded from this analysis. Four-year-olds were at chance (26 out of 48 four-year-olds, binomial 1/2, n.s.).

Discussion

The current study found that when four-year-olds’ WM and ID were jointly taxed, four-year-olds’ performance was similar to three-year-olds’ performance in high ID false belief tasks. More specifically, four-year-olds frequently attributed a true belief to all of the agents and were at chance when their responses were categorized as either attributing mostly false beliefs or attributing mostly true beliefs. This finding is perhaps not surprising considering research on adults’ false belief performance when EF is taxed. For example, McKinnon and Moscovitch (2007) found that individuals with smaller WM capacity (elderly participants) have more difficulty reasoning about an agent’s mental states than individuals with a larger WM capacity (undergraduates). In addition, researchers have found that when undergraduates engage in a simultaneous EF task (inhibition task, see Qureshi, Apperly, & Samson, 2010 and WM task, see McKinnon & Moscovitch, 2007), undergraduates demonstrated similar perspective taking failures.

General Discussion for Experiments 5-7

Three-year-olds’ WM capacity for tracking agents with distinct false beliefs is limited to approximately 3 agents while four-year-olds’ WM capacity is limited to approximately 5 agents. we found that when preschoolers’ WM capacity is taxed, their ability to attribute a false belief to the agents in the task does not deteriorate dramatically. Instead, there is a gradual increase of “WM” errors, where both three- and four-year-olds attribute the incorrect false belief to the given agent. Taken together, these findings demonstrate that preschoolers’ WM capacity for tracking agents and their false beliefs
echo previous work in adults’ WM capacity for agent tracking in conversations and in social interactions illustrated in various media forms (Cohen, 1971; Matthews & Barrett, 2005; Krems & Dunbar, 2013; Stiller et al., 2005).

When both WM and ID were taxed together, four-year-olds’ performance shares similar characteristics to three-year-olds’ performance as revealed by careful error analysis. These findings favor the expression view, which suggests that children’s false belief performance is subject to EF task demands, and provide evidence against the emergence view, which suggests that children’s false belief performance should be stable once they have acquired sufficient EF to allow the concept of false belief to emerge.

Another EF-Theory of Mind account, the “competence” view, argues that that theory of mind is intimately tied with EF (Apperly, Samson, & Humphreys, 2009; Frye, Zelazo, & Palfai, 1995; Russell, 1996). Namely, in order to reason about others’ mental representations of the world, i.e., use theory of mind, one must use EF, presumably WM and “other” EF factors, to construct and maintain multiple representations. The findings in the studies presented are partially consistent with this view. The studies so far have demonstrated preschoolers’ capacity for tracking multiple agents with distinct false beliefs, suggesting that the inability to track multiple representations of the world cannot be the cause of failure. Instead, failure appears to be mainly caused by ID and WM only when WM capacity nears its limit. The current studies, however, are unable to disentangle theory of mind and EF completely, which leaves room for the “competence” view to continue to be a possibility for describing how EF affects theory of mind.
Chapter 5

The purpose of the current chapter is to report a comparison across Experiments 1-7. The main question addressed in Chapters 2-4 is how executive function, particularly inhibition and working memory, influences preschoolers’ false belief performance. Although Chapters 2-4 report which EF skill plays a larger role in false belief performance and how EF relates to the development of theory of mind reasoning, one question still remains: what is the nature of preschoolers’ errors in the verbal change location false belief task? In the following sections, we first address preschoolers’ errors by reviewing how the multiple agent false belief tasks deviated from the standard 50% baseline. Next, we discuss the most common error that preschoolers made, the all true belief error. Finally, we describe the true belief responses in the low demand task. We conclude the chapter with a general discussion and conclusions.

Deviating from the standard 50% baseline

In Experiments 1-7, we followed the common practice in the theory of mind literature of assuming a uniform distribution model for H₀ against which to test the observed data. In the standard Sally and Anne task (which asks only about Sally), this dictates a chance baseline of .5 for passing and failing. We extended this model to multiple agents calculating the baseline probability of a “fully correct” response as 1/Beliefs^Agents. We found the uniform model fitted poorly with our data: (1) we were able to reject H₀ repeatedly using alpha levels that were many orders of magnitude smaller than hitherto with sample size n=50; (2) our data revealed that, despite the strong effects for passing responses noted in (1), the true belief attribution error remains attractive, at nearly the same effect size and even for four-year-olds, with multiple agents.
Rather than assuming a uniform distribution in this chapter we will examine a model based on observed probabilities. It is well-established that preschoolers’ passing performance on a standard single agent false belief task has a probability around .3 for three-year-olds and .7 for four-year-olds. Therefore, a simple model for success (defined as being “fully correct”) for a multiple agent false belief task is the combination of these probabilities for each agent, \( P(\text{success})^A \). Put differently, to derive the probability of being fully correct on a multiple agent task, we use the probability of success on the single agent false belief task raised to the power of the number of questions the child is asked in a multiple agent task. For example, in the Double Agent High Demand task, the three-year-olds’ probability of being “fully correct” is: \( P(\text{success})^A = .3^2 = .09 \). We calculated the predicted probabilities for each level of agents and demand. Existing findings for Low Demand tasks are less abundant. We used data from Kikuno et al. (2007), Zaitchik (1991), and unpublished data from our lab (collected by Dr. Lu Wang); these are all studies using change of location tasks where the final location is uncertain and non-specific to ensure similarity to the tasks used here. Across these studies, 52 out of 82 three-year-olds passed, \( P = .634 \). Data on low demand performance in 4-year-olds is even scarcer: we used data from Kikuno et al. (2007), Wellman and Bartsch (1988), and Zaitchik (1991): 68 out of 84 four-year-olds passed, \( P = .81 \).

Tables 1 and 2 show the predicted success rates based on findings in the existing literature on the single agent false belief task under “Predicted” columns and provide corresponding success rates we observed in Experiments 1-7 under the “Observed” columns.

---

9 Here “P” refers to probability.
Table 1. Three-year-olds’ Probabilities of Success: Predicted vs. Observed Proportions

<table>
<thead>
<tr>
<th>Agents</th>
<th>LD Predicted</th>
<th>LD Observed</th>
<th>HD Predicted</th>
<th>HD Observed</th>
</tr>
</thead>
<tbody>
<tr>
<td>Single</td>
<td>.634</td>
<td>.300</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Double</td>
<td>.402</td>
<td>.400</td>
<td>.090</td>
<td>.200</td>
</tr>
<tr>
<td>Triple</td>
<td>.255</td>
<td>.400</td>
<td>.030</td>
<td>.100</td>
</tr>
<tr>
<td>Quadruple</td>
<td>.162</td>
<td>.220</td>
<td>.001</td>
<td>not tested</td>
</tr>
</tbody>
</table>

Table 2. Four-Year-Olds’ Probabilities of Success: Predicted vs. Observed Proportions

<table>
<thead>
<tr>
<th>Agents</th>
<th>LD Predicted*</th>
<th>LD Observed</th>
<th>HD Predicted</th>
<th>HD Observed</th>
</tr>
</thead>
<tbody>
<tr>
<td>Single</td>
<td>.810</td>
<td>.7</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Double</td>
<td>.656</td>
<td>.640</td>
<td>.490</td>
<td>.52</td>
</tr>
<tr>
<td>Triple</td>
<td>.531</td>
<td>.560</td>
<td>.343</td>
<td>.34</td>
</tr>
<tr>
<td>Quadruple</td>
<td>.430</td>
<td>.500</td>
<td>.240</td>
<td>.16</td>
</tr>
<tr>
<td>Pentuple</td>
<td>.349</td>
<td>.180</td>
<td>.168</td>
<td>not tested</td>
</tr>
</tbody>
</table>

* Four-year-olds’ performance on the single agent low demand task is extrapolated from tasks considered low demand in the general literature.

Figure 30 illustrates the similarity between the observed probabilities for success (defined as “fully correct”) and the predicted probabilities based on single agent probabilities drawn from the literature and project “forwards” to the multiple agent by demand tasks. We refer to this way of predicting from existing data to our experiments, the “Forwards” predictions of the model. It is immediately clear that this model fits better than the uniform model. Both the predicted and the empirical data lines suggest that the probability for success decreases as the number of agents increases. However, it is consistent with the probability of success for each agent/question remaining constant (that is, the probability for success in the corresponding single agent task).
Figure 30. Predicted Probabilities of Success vs. Observed Proportions: A comparison between the predicted probabilities of success derived from preschoolers’ performance on the verbal Sally-Anne task and the observed proportions derived from Experiments 1-7. The top panels illustrate three-year-olds and the bottom panels illustrate four-year-olds.

We calculated the probability of failure in the same way as the probability of success except we used the rate of failure for each age group and demand level. For example, in the High Demand tasks, we used .7 failure for three-year-olds and .3 failure for four-year-olds. Using preschoolers’ failure rate, we computed the probability for failure using the “Forwards” predictions of the model, $P(\text{failure})^A$. The predicted baseline probabilities for failure based on the single agent false belief task were compared to preschoolers “failure” performance (defined as “all true belief”) in Experiments 1-7 (shown in Tables 3 and 4). Again, the predicted failure rates based on findings in the
existing literature on the single agent false belief task under “Predicted” columns and provide corresponding failure rates we observed in Experiments 1-7 under the “Observed” columns.

**Table 3. Three-year-olds’ Probabilities of Failure: Predicted vs. Observed Proportions**

<table>
<thead>
<tr>
<th>Agents</th>
<th>LD Predicted</th>
<th>LD Observed</th>
<th>HD Predicted</th>
<th>HD Observed</th>
</tr>
</thead>
<tbody>
<tr>
<td>Single</td>
<td>.366</td>
<td>.700</td>
<td>.490</td>
<td>.440</td>
</tr>
<tr>
<td>Double</td>
<td>.134</td>
<td>.420</td>
<td>.343</td>
<td>.380</td>
</tr>
<tr>
<td>Triple</td>
<td>.049</td>
<td>.300</td>
<td>.380</td>
<td>.380</td>
</tr>
<tr>
<td>Quadruple</td>
<td>.018</td>
<td>.380</td>
<td>.240</td>
<td>not tested</td>
</tr>
</tbody>
</table>

**Table 4. Four-Year-Olds’ Probabilities of Failure: Predicted vs. Observed Proportions**

<table>
<thead>
<tr>
<th>Agents</th>
<th>LD Predicted*</th>
<th>LD Observed</th>
<th>HD Predicted</th>
<th>HD Observed</th>
</tr>
</thead>
<tbody>
<tr>
<td>Single</td>
<td>.190</td>
<td>.300</td>
<td>.90</td>
<td>.280</td>
</tr>
<tr>
<td>Double</td>
<td>.036</td>
<td>.240</td>
<td>.027</td>
<td>.240</td>
</tr>
<tr>
<td>Triple</td>
<td>.001</td>
<td>.320</td>
<td>.008</td>
<td>.440</td>
</tr>
<tr>
<td>Quadruple</td>
<td>.0002</td>
<td>.360</td>
<td>.002</td>
<td>not tested</td>
</tr>
<tr>
<td>Pentuple</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

* Four-year-olds’ performance on the single agent low demand task is extrapolated from tasks considered low demand in the general literature.

Figure 31 illustrates the observed probabilities for failure (defined as “all true belief”) and the predicted probabilities based on single agent probabilities. Unlike the predicted values for “fully correct” responses, the “Forwards” predictions for failure greatly underestimates the failure rate for both three- and four-year-olds in the low demand tasks. In addition, the “Forwards” prediction also underestimates the failure rate for four-year-olds in the high demand task. The “Forwards Model” only closely predicts three-year-olds’ failures on the high demand tasks. Although the “Forwards” predictions fits fairly closely to preschoolers’ observed “fully correct” performance, it does a strikingly poor job of predicting failure.
We now turn to an alternative way of predicting performance that we shall call “Backwards” predictions of the model. Unlike “Forwards” predictions, which project from the single agent task “forwards” to the multiple agent tasks, the “Backwards” prediction projects from the multiple agent tasks “backwards” to the single agent task. Using the backwards model, we can estimate the probability for a successful/unsuccesful response in the single agent task by using the observed proportions of “fully correct” responses and “all true belief” errors obtained from the multiple agent tasks. Using the observed proportions, we derive the probability for success/failure on a single agent task.
by taking the nth root of the observed proportion, where “n” is the number of questions the child is asked in a multiple agent task. For example, in the Double Agent High Demand task, we took 2nd root (square root) of four-year-olds’ proportion of “fully correct” responses (.52), \( \sqrt{.52} \) to obtain a P=.721. This P indicates four-year-olds’ estimated probability for success on the single agent task.

Tables 5-8 display the observed “All Correct” and “All True Belief” proportions as well as the estimated probabilities for success and failure for the single agent. We took the average of each age group’s estimate of success per demand and found that the estimates converge with the single agent literature. For example, the average estimate of four-year-olds’ success in the single agent high demand task is .681. This is strikingly similar to the traditional finding that approximately “70% of four-year-olds pass” the single agent task. Like the “Forwards” predictions the estimated probabilities for success derived by the “Backwards” estimations are consistent with the probability of success for each agent remaining constant. In contrast, when we examine the average estimated probability for failure in each age group per demand level, the “Backwards” estimations (like the “Forwards” predictions) does a poor job in estimating failure for the single agent task. Contrary to the “Forwards” predictions, which vastly underestimated error, the “Backwards” estimations greatly overestimated the error rate for the single agent task for both three- and four-year-olds on the low demand tasks. In addition, the “Backwards” estimations also greatly overestimated the error rate for four-year-olds on the high demand task. Like the “Forwards” predictions, the “Backwards” estimations closely estimated three-year-olds’ failure rate on the single-agent high-demand task \( (M = .694) \).
Table 5. Estimated Probabilities for the Single Agent Task: Three-year-olds on the Low Demand Tasks

<table>
<thead>
<tr>
<th>Tasks</th>
<th>Estimated Success</th>
<th>Observed “Fully Correct”</th>
<th>Estimated Failure</th>
<th>Observed “All True Belief”</th>
</tr>
</thead>
<tbody>
<tr>
<td>Double</td>
<td>.632</td>
<td>.400</td>
<td>.648</td>
<td>.420</td>
</tr>
<tr>
<td>Triple</td>
<td>.737</td>
<td>.400</td>
<td>.669</td>
<td>.300</td>
</tr>
<tr>
<td>Quadruple</td>
<td>.685</td>
<td>.220</td>
<td>.785</td>
<td>.380</td>
</tr>
<tr>
<td>Average Estimate</td>
<td>.685</td>
<td></td>
<td></td>
<td>.701</td>
</tr>
</tbody>
</table>

Table 6. Estimated Probabilities for the Single Agent Task: Four-year-olds on the Low Demand Tasks

<table>
<thead>
<tr>
<th>Tasks</th>
<th>Estimated Success</th>
<th>Observed “Fully Correct”</th>
<th>Estimated Failure</th>
<th>Observed “All True Belief”</th>
</tr>
</thead>
<tbody>
<tr>
<td>Double</td>
<td>.800</td>
<td>.640</td>
<td>.490</td>
<td>.240</td>
</tr>
<tr>
<td>Triple</td>
<td>.824</td>
<td>.560</td>
<td>.604</td>
<td>.220</td>
</tr>
<tr>
<td>Quadruple</td>
<td>.841</td>
<td>.500</td>
<td>.752</td>
<td>.320</td>
</tr>
<tr>
<td>Pentuple</td>
<td>.710</td>
<td>.180</td>
<td>.815</td>
<td>.360</td>
</tr>
<tr>
<td>Average Estimate</td>
<td>.794</td>
<td></td>
<td></td>
<td>.665</td>
</tr>
</tbody>
</table>

Table 7. Estimated Probabilities for Single Agent: Three-year-olds on the High Demand Tasks

<table>
<thead>
<tr>
<th>Tasks</th>
<th>Estimated Success</th>
<th>Observed “Fully Correct”</th>
<th>Estimated Failure</th>
<th>Observed “All True Belief”</th>
</tr>
</thead>
<tbody>
<tr>
<td>Double</td>
<td>.447</td>
<td>.200</td>
<td>.663</td>
<td>.440</td>
</tr>
<tr>
<td>Triple</td>
<td>.464</td>
<td>.100</td>
<td>.724</td>
<td>.380</td>
</tr>
<tr>
<td>Average Estimate</td>
<td>.456</td>
<td></td>
<td></td>
<td>.694</td>
</tr>
</tbody>
</table>

Table 8. Estimated Probabilities for Single Agent: Four-year-olds on the High Demand Tasks

<table>
<thead>
<tr>
<th>Tasks</th>
<th>Estimated Success</th>
<th>Observed “Fully Correct”</th>
<th>Estimated Failure</th>
<th>Observed “All True Belief”</th>
</tr>
</thead>
<tbody>
<tr>
<td>Double</td>
<td>.721</td>
<td>.520</td>
<td>.529</td>
<td>.280</td>
</tr>
<tr>
<td>Triple</td>
<td>.698</td>
<td>.340</td>
<td>.621</td>
<td>.240</td>
</tr>
<tr>
<td>Quadruple</td>
<td>.632</td>
<td>.160</td>
<td>.814</td>
<td>.440</td>
</tr>
<tr>
<td>Average Estimate</td>
<td>.681</td>
<td></td>
<td></td>
<td>.655</td>
</tr>
</tbody>
</table>
The All True Belief Error

Across all seven experiments, preschoolers had the opportunity to make a variety of errors (up to 7,775 for Pentuple Agent!), but the majority of preschoolers who made errors were attracted to making the All True Belief Error. This primary error was made by both age groups regardless of the number of agents and the demand type (see Figure 32).

Figure 32. Proportion of Three- and Four-year-olds’ All True Belief Errors in Experiments 1-7.

When compared to other types of error that were made, the All True Belief Error is unique because it involves the same belief attribution for all of the agents. Apart from the All True Belief Error, attributing the same belief to all of the agents was extremely uncommon. In Experiments 1-4, preschoolers rarely attributed the same false belief to
two agents, let alone the same false belief to all the agents. For Experiments 1-4 (N=400), only about 2% of preschoolers attributed the same false belief to all agents. Attributing the same false belief to more than one agent became more common when WM load was high (albeit not as high as the All True Belief Error). Twenty percent of three-year-olds in the Four Agent Low Demand task and 22% of four-year-olds in the Five Agent Low Demand task attributed the same false belief to at least two agents. Further, 10% of three-year-olds and 10% of four-year-olds attributed the same false belief (the last agent’s false belief) to all of the agents in the Four Agent Low Demand and Five Agent Low Demand task respectively. See Figure 33 below for the individual responses in Experiments 1-7.
Figure 33. Individual Responses for Experiments 1-7.
Attributing a True Belief to an Agent in the Low Demand Tasks

In the standard single agent task, preschoolers, who make an incorrect response, systematically point to where they (the subjects) believe the desired target is located. Because of this systematic error, some argue that preschoolers are able to represent another individual’s mental state, but they make this error because their own belief is more salient than Sally’s false belief (Friedman & Leslie, 2005; Leslie & Thaiss, 1992; Roth & Leslie, 1998). To reduce the saliency of the preschoolers’ own belief, researchers have modified the standard false belief task by making the desired target’s final location “vague.” This manipulation undermines preschoolers’ certainty about their own belief and thus reduces the saliency of their own belief and making the selection between Sally’s false belief and their own belief easier (Carpenter et al., 2002; Southgate et al., 2007; Wang & Leslie, 2016). Although removing the desired target to an indefinite location helps preschoolers succeed on the false belief task, it does not completely remove their inclination to attribute the true belief. In the low demand tasks presented in this dissertation, preschoolers still attributed true beliefs to the agents. These true belief attributions included “vague” type of responses such as reporting that the agent will not find the desired target or “I don’t know.” Additionally, preschoolers also spontaneously generated their own belief of the desired target’s location and predicted the agent will search there. Table 9 includes all the categories that were coded as true belief attributions and examples for each category.

We used the same coding scheme for children’s responses in the High Demand tasks and found that only a small percentage of responses in these tasks fit the description of the categories in Table 9. Across all High Demand tasks (N=250), only nine
preschoolers (4 three-year-olds and 5 four-year-olds) made responses that could be categorized as “vague” true belief attributions. Six of these preschoolers predicted that the agent would search in multiple locations (Category 2) while the other three indicated that a search would result in not finding the desired target (Category 1). For the Low Demand tasks, 132 preschoolers (61 three-year-olds and 71 four-year-olds) made “vague” true belief responses. We coded each true belief attribution as one of the 11 categories. The counts are illustrated in Figure 34.

Table 9. Examples of Responses Coded as “True Belief” Responses in the Low Demand Tasks.

<table>
<thead>
<tr>
<th>Category</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>C1: Indicating a search will result in not finding the target</td>
<td>No/Nowhere/He/she won’t find it</td>
</tr>
<tr>
<td>C2: Reporting “I don’t know”</td>
<td>I don’t know</td>
</tr>
<tr>
<td>C3: Generating a possible location for target</td>
<td>Up in the sky/In the woods/On the farm/In the tree/Florida/Under the table/Away from her home/Inside/Forest/Playroom/Pool/Water/Pool/In a garden/At a beach/Bushes/In the air/Nest/Woods/Inside house/At the farm/In toy box/Inside school/In a jungle/Under the carpet/Kitchen table</td>
</tr>
<tr>
<td>C3: Predicting search in multiple locations</td>
<td>Everywhere/Pointing to multiple locations/All over the place</td>
</tr>
<tr>
<td>C5: Repeating the target’s last known location</td>
<td>Far far away/Far away/Away</td>
</tr>
<tr>
<td>C6: Predicting a search outside of the room</td>
<td>Outside/Pointing outside/Out there/Out of room</td>
</tr>
<tr>
<td>C7: Predicting a search for the cat (Double Low Demand)</td>
<td>Cat/In the cat’s mouth/The cat has it/The cat’s tummy/From the cat</td>
</tr>
<tr>
<td>C8: Repeating the target’s last actions</td>
<td>Hopped away/Flew away/Flying</td>
</tr>
<tr>
<td>C9: Indicating the target is gone</td>
<td>Cookie/Frog/Chick will be gone</td>
</tr>
<tr>
<td>C10: Reporting that the agent does not know/see</td>
<td>He/she wouldn’t know/He/she didn’t see</td>
</tr>
<tr>
<td>C11: Uncodeable</td>
<td>Pointed to agent’s foot/Yeah</td>
</tr>
</tbody>
</table>
We investigated the nature of preschoolers’ errors by examining their errors in the multiple agent false belief tasks. In a multiple agent task, the number of responses exponentially increases as the number of agents with distinct false beliefs increases. Despite the number of possible responses, the majority of preschoolers in both age groups responded in two different ways: attributing the correct false belief given the agent or attributing a true belief to all the agents. Because preschoolers demonstrated a tendency to respond in these two ways, we decided to deviate from comparing preschoolers’
performance to a uniform distribution and instead examined performance using two methods of predicting the proportion of success and failure.

The “Forwards” method used the standard single agent false belief task to predict the probability of success and failure for each multiple agent false belief task. The “Backwards” method used the observed results from the multiple agent false belief task to predict the probability of success and the probability of failure on the single agent task. The “Forwards” method closely predicted preschoolers’ success rate across number of agents and demand, but underestimated the four-year-olds’ errors in the low and high demand tasks as well as three-year-olds’ errors in the low demand task. The “Backwards” method also accurately estimated preschoolers’ “success” performance for the single agent task across number of agents and demand. But this method overestimated four-year-olds’ errors in the low and high demand tasks as well as three-year-olds’ errors in the low demand task. These methods accurately predicted/estimated three-year-olds’ errors on the high demand tasks. Because the multiple agent false belief tasks are novel tasks, it is reassuring that these methods reveal that performance from the standard single agent false belief task (a highly replicated result) can predict success on the multiple agent tasks and the multiple agent false belief task can estimate success on the single agent false belief task. It is unclear why both methods were only able to predict/estimate three-year-olds’ errors on high demand tasks and not the others (threes on low demand tasks, fours on both low and high demand tasks). This raises questions about preschoolers’ failures, which can be addressed in future research.

Out of all the errors, the All True Belief Error was the most frequent and continued to be the largest proportion of the errors committed regardless of the number of
agents and the demand type. This type of response is unique because it includes the same belief attribution for all the agents. It is not until preschoolers reach their WM capacity limit (four agents for three-year-olds and five agents for four-year-olds) that they begin to attribute the same false belief to all the agents. When they reached their capacity for tracking agents, a handful of the preschoolers attributed the last agent’s false belief to all the agents. Although this is notable, it is still not as frequent as the All True Belief Error.

It is unclear why attributing a true belief to all the agents is prevalent across both age groups and all tasks. It could be due to inhibitory control development. Namely, we can categorize the majority of the preschoolers from Experiments 1-7 into three categories: 1) preschoolers who make All True Belief Errors, 2) preschoolers who make bleeding true belief errors (attributing a true belief for the first agent then a false belief to the remaining), and 3) preschoolers who only make false belief attributions. First, preschoolers, who make an All True Belief Error, could be those whose inhibition is least developed. Their underdeveloped inhibition may cause them to be unable to inhibit the prepotent true belief response despite multiple opportunities to attribute correctly a false belief. Second, those who make bleeding true belief errors could be those whose inhibition is bordering the threshold required to inhibit the prepotent true belief response. Having multiple opportunities buys time for the preschooler to inhibit the salient true belief response. Third, children who only attribute false beliefs to the agents could be those whose inhibition meets the threshold required to inhibit the prepotent true belief response. This categorization of children, however, is speculative. Future studies can investigate this further by examining the children’s inhibitory control on a separate task and relating their performance to their performance in a multiple agent false belief task.
General Discussion for Chapters 1-7

Experiments 1-7 shed light on three aspects of preschoolers’ verbal false belief performance. First, three- and four-year-olds have the capacity to track multiple agents: approximately three to four agents for three-year-olds and approximately four to five agents for four-year-olds. This finding suggests that WM does not limit three-year-olds’ performance in the single agent false belief task, but only plays a minor role. It is not until children are required to track near or above their WM capacity that preschoolers’ performance suffers. Instead, Experiments 1-7 showed that three-year-olds’ performance in the single agent false belief task is limited by the ID. When ID was high, three-year-olds typically erred by attributing a true belief to all of the agents. When ID was low, however, three-year-olds attributed the correct false belief to the given agent. Further, ID also affected four-year-olds’ performance, but only when it was coupled with high WM demand. More specifically, when four-year-olds were required to track four agents with distinct false beliefs in a high ID task, their performance shared a strong resemblance to three-year-olds’ performance in high ID false belief tasks. Furthermore, as noted on pp. 75-76, when demand is held constant at the high level and four-year-old performance on Triple agent is compared with that on Quadruple agent, fours perform significantly worse with four agents than with three, showing fewer All Correct and more All True Belief responses, while Binding and Mixed errors are unchanged.

Second, the findings in Experiments 1-7 provide evidence against the emergence view and in favor for the expression view. The former account suggests that EF is necessary to engender the concept of belief while the latter account claims that EF is unnecessary to represent false beliefs, but it is necessary to engage in a verbal false belief
task. Experiments 1-7 revealed that when inhibition was lowered, three-year-olds’ performance improved. More striking, when four-year-olds’ EF was heavily taxed (high WM and high ID), their performance resembled three-year-olds’ false belief performance in high ID tasks. According to the emergence view, four-year-olds should never perform like three-year-olds because they have already developed the necessary EF skills to entertain another’s perspective, which allows them to pass a verbal false belief task. Taken together, these findings suggest that EF continues to be an influential factor in verbal false belief performance.

Third, testing preschoolers’ multiple agent theory of mind allowed the first examination of preschoolers’ errors in a false belief task. Across seven multiple agent false belief tasks, where the number of possible responses varied, the two most common responses were: 1) attributing the correct false belief to the given agent (i.e., being “fully correct”) and 2) attributing a true belief to all of the agents (i.e., making an All True Belief Error). This was regardless of the number of agents and the demand type. The All True Belief Error was especially unique because unlike the other responses observed, the All True Belief Error involved the same belief attribution to all agents. It is unclear why preschoolers were prone to attribute the same (true) belief to all of the agents, but not the same false belief to all of the agents. It is possible that the preschoolers who attributed a true belief to all agents were those whose inhibition was too immature to overcome the saliency of the true belief, hence why they attributed the true belief to all agents. Because children’s inhibition was not assessed, we are unable to verify this ad-hoc explanation. Future studies can take an independent measure of inhibition and examine whether
preschoolers, who attribute the true belief to all the agents, score lower on inhibition tasks.

There were some limitations to the current series of experiments. We found that ID greatly affected preschoolers’ false belief performance while WM only minimally affected preschoolers’ false belief performance. One limitation of the series of multiple agent false belief tasks was how we manipulated inhibition. In the current studies, inhibition was tested as a binary measure (either high or low). But, like WM, inhibition is not considered to be a binary measure, but rather a graded measure. In Experiments 2 (Double Agent Low Demand) and 3 (Triple Agent Low Demand), we observed that preschoolers made qualitatively more “mostly false belief” responses in Experiment 3 than Experiment 2. As discussed in Chapter 2, this may be due to our manipulation of the child’s knowledge of the desired target’s final location. In Experiment 2, another character took the desired target to an unknown location, but in Experiment 3, the desired target disappeared. For the former manipulation, the desired target’s final location was somewhat uncertain. This is because although the final location was uncertain, it was clear who had the desired target at the end of the story. In contrast, for the latter manipulation, preschoolers were told that the desired target “disappeared,” making the desired target’s final location completely uncertain. Future studies can assess if preschoolers’ responses can be modulated by the child’s uncertainty of the desired target’s final location by incrementing the certainty of the target’s final location. Another limitation was the cross-sectional design. Although a cross-sectional design has its advantages, a longitudinal design would allow researchers to examine how WM capacity for agents develops and how preschoolers’ responses (and errors!) change over time.
Future studies can further examine how EF affects other areas of theory of mind. For example, some argue there is an additional conceptual theory of mind milestone achieved around the age of 6-7 years (Perner, & Howes, 1992; Perner, & Wimmer, 1995). This later theory of mind milestone is referred to as second order theory of mind. In contrast to first order theory of mind, where Sally (falsely) believes that the marble is in the basket, second order theory of mind is the ability to reason that Anne (falsely) believes that Sally believes that the marble is in the box. While some believe that six-year-olds are making a conceptual breakthrough (Perner & Howes, 1992; Perner & Wimmer, 1995), others argue that the additional lag between first order theory of mind and second order theory of mind is due to children’s developing WM capacity, i.e., the story’s complexity, length, and number of characters (Sullivan, Zaitchik, & Tager-Flusberg, 1994). Because the current studies show that three- and four-year-olds can track at least three agents with distinct false beliefs, it is unlikely that the number of characters in a second order theory of mind task (2-3 characters) is a limiting factor. Why second order theory of mind is challenging for four-year-olds still remains unknown and can be explored in future studies.

In the multiple agent false belief tasks, preschoolers were required to track multiple agents and their distinct false beliefs simultaneously, but the cognitive mechanism for tracking multiple agents has yet to be determined. Currently, there are several models describing how individuals calculate a single agent’s false belief (see Baker, Saxe, and Tenebaum, 2011; Leslie & Polizzi, 1998; Goodman et al., 2006; Wang, Hemmer, & Leslie, under revision), but for multiple agents, it is unknown whether
individuals sequentially compute each agent’s false belief or simultaneously compute all of the agents’ false beliefs.

Conclusions

The findings reported in this dissertation provided the first empirical evidence that preschoolers are capable of tracking a small group of individuals with distinct false beliefs. Preschoolers’ capacity appears to be limited to approximately 3-4 agents, making it plausible that this capacity may have been adapted to track a small “family.” In addition, we demonstrated that for each additional agent added (i.e., increasing WM demand), the cognitive load only appears to increase slightly. In contrast, when ID was high, the cognitive load appears to increase dramatically, suggesting that ID plays a major role in preschoolers’ performance in the standard single agent false belief task while WM demand (prior to reaching to 3-4 agents) only plays a minor role. Finally, although analyzing error patterns plays a large role in studying underlying cognitive processes, this data was previously unavailable to theory of mind researchers because the standard single agent false belief task only permits one way of failing. Because there has only ever been one way of failing, error patterns in false belief tasks remained unexplored. The current dissertation provides a new method for not only testing preschoolers’ multiple agent theory of mind, but also allows preschoolers to respond in 9, 64, 625, and 7,776 ways! Despite having a variety of possible responses, preschoolers in all seven experiments tended to respond in two ways: attributing the correct false belief to each agent and attributing to all agents the true belief. It is unclear why these two responses are the most prominent, but this is a question that can be explored for future research.
References


Surian, L., & Geraci, A. (2011). Where will the triangle look for it? Attributing false beliefs to a geometric shape at 17 months. *British Journal of Developmental Psychology, 30*(1), 30-44.


