# Children's Application of Theory of Mind in Reasoning and Language

LIESBETH FLOBBE[1], RINEKE VERBRUGGE[2], PETRA HENDRIKS[3], and IRENE KRÄMER[4]

[1]*E-mail: L.Flobbe@ai.rug.nl*

[2]*Institute of Artificial Intelligence, University of Groningen, Bernoulliborg, room 355, P.O. Box 407, 9700 AK Groningen, The Netherlands*

*E-mail: rineke@ai.rug.nl*

[3]*Center for Language and Cognition Groningen (CLCG), University of Groningen, P.O. Box 716, 9700 AS Groningen, The Netherlands*

*E-mail: P.Hendriks@rug.nl*

[4]*Centre for Language Studies (CLS), Radboud Universiteit Nijmegen, P.O. Box 9103, 6500 HD Nijmegen, The Netherlands, and Iselinge School of Education, P.O. Box 277, 7000 AG Doetinchem, The Netherlands*

*E-mail: I.Kramer@let.ru.nl*

**Abstract.** Many social situations require a mental model of the knowledge, beliefs, goals, and intentions of others: a Theory of Mind (ToM). If a person can reason about other people's beliefs about his own beliefs or intentions, he is demonstrating second-order ToM reasoning. A standard task to test second-order ToM reasoning is the second-order false belief task. A different approach to investigating ToM reasoning is through its application in a strategic game. Another task that is believed to involve the application of second-order ToM is the comprehension of sentences that the hearer can only understand by considering the speaker's alternatives. In this study we tested 40 children between 8 and 10 years old and 27 adult controls on (adaptations of) the three tasks mentioned above: the false belief task, a strategic game, and a sentence comprehension task. The results show interesting differences between adults and children, between the three tasks, and between this study and previous research.

**Keywords:** false belief, second-order reasoning, sentence comprehension, strategic game, Theory of Mind

## 1. Introduction

### 1.1 Theory of Mind

Many everyday reasoning tasks require reasoning about the knowledge and intentions of other people. The capacity for this kind of reasoning is sometimes called mind reading. A common approach to studying this capacity uses the phrase 'theory of mind' (ToM), first coined in the article "Does the chimpanzee have a theory of mind?" (Premack and Woodruff, 1978). In the ToM approach a child's cognitive development is understood by assuming that the child acquires a 'theory of mind': a mental model of the social world similar to folk psychology. A child who has a theory of mind understands that other people have minds too, with beliefs, desires, and intentions possibly distinct from his own. He can formulate hypotheses about what those beliefs, desires, and intentions are.

While much research has focused on very early development of Theory of Mind, the focus of the present study is on second-order Theory of Mind, which develops later than first-order ToM. ToM reasoning can be classified by its order of mental state attribution. Reasoning about other people's beliefs and intentions about simple world facts is first-order reasoning. Examples of first-order attributions are: "Mary believes that the ball is in the bag" or "You intend to take the left cup". However, if a person takes into account the other person's beliefs and intentions about the minds of others (including the first person's), that person uses second-order reasoning. Examples of second-order mental state attributions are: "Mary believes that John believes that the ball is in the closet" or "You believe that I believe that the box contains a pencil". Thus, the famous false-belief task about Maxi and his mother tests for first-order mental state attributions: Does the child correctly conclude that Maxi will look for his chocolate in its original location, while the child knows that Maxi's mother displaced it while Maxi was gone, thus attributing a false belief to Maxi (Wimmer and Perner, 1983)?

It is the aim of the present article to contribute to charting the late development of second-order ToM by investigating participants' performance on tasks in three different domains – a strategic game, a grammatical task, and a standard

second-order false belief task. Successful performance on each of these tasks requires the application of a second-order Theory of Mind.

The article addresses two main issues: the developmental discrepancy of first and second-order ToM, and the task dependence of ToM. As to the first of these issues, children generally pass the standard false belief task by age 4, but it takes another two years for them to pass a similar task if it requires second-order ToM (Tager-Flusberg and Sullivan, 1994; see also Steerneman et. al., 2003). A study by Hedden and Zhang (2002) suggests a cause for this difference, namely that the processing of second-order ToM is more costly than that of first-order ToM. Hedden and Zhang's task was a strategic game, very different from a standard false belief task, and it only included adult participants. Whereas participants were generally good at applying first-order reasoning when the game so required, second-order ToM reasoning was seriously flawed with most of the adults. This study suggests inherent difficulty with second-order ToM reasoning, which may be responsible for the delay in the surfacing of second-order ToM in child development. If this is the case, we should see that children have less difficulty in applying first-order rather than second-order ToM, not only in a false belief task, but also in a game task, and that children perform worse than adults on such a task. Thus, we made the game task the focus of our investigations.

This brings us to the question of whether we expect differential performance on tasks involving different cognitive domains. Many studies focus on the question of whether individuals have a Theory of Mind. The task of experimental research is, then, to find a way to tap into this mental ability while avoiding its being masked by performance factors caused by a given experimental task. However, the lower boundaries of ToM manifestations have been pushed down to increasingly lower ages, and the upper boundaries of failed ToM performance may need to be lifted entirely, as it becomes clear that even adults do not display perfect ToM performance. Thus, the question of which conditions promote or hamper the use of ToM, and why, has steadily gained importance (see, for example, De Villiers, 2007). The present article's second aim is to contribute findings concerning second-order ToM to this discussion, explicitly comparing the results from different experimental tasks. The next section briefly sketches the background of this discussion.

### 1.2 Task dependence

Research on ToM development shows that whether participants successfully apply Theory of Mind strongly depends on the task, a most striking example of which we find in the discrepancies between the standard first-order false belief task (Wimmer and Perner, 1983) and a recent looking time experiment by Onishi and Baillargeon (2005). In the standard first-order false belief task, the child is asked to predict the behaviour of another person, for example where the person will search for an object. To make a correct prediction the child must understand that this person holds a false belief that is different from the child's own, true, beliefs. Success at such a task indicates clearly that the child knows that other people have beliefs, and that the child can distinguish between its own beliefs and those of others. Children at age 3 still fail first-order false belief tasks, but children at age 4 or older pass them. In the study by Onishi and Baillargeon, the dependent variable is looking time. Fifteen-month-old children were shown to distinguish between cases in which an actor looked in a place in which the actor knew the object that she looked for was not to be found, and cases in which the actor looked in the right place. DeVilliers (2007) points out that the vast discrepancy between 15 months at Onishi and Baillargeon's task and the passing age for the first-order false belief task may well lie in the task demands, in that the latter task, but not the former, requires decision making.

Regarding studies such as Onishi and Baillargeon (2005) that claim to show very early presence of ToM, questions have been raised as to whether correct performance on the tasks really requires ToM (see Perner and Ruffman, 2005). The limitations posed by the cognitive and communicative development of young children put severe restrictions on the format of experimental tasks. Therefore, to compare the application of ToM across different tasks, the study of later ToM development may be particularly suited. The work of Keysar and colleagues provides examples of how adults do not always correctly draw upon first-order ToM. Keysar, Lin, and Barr (2003) report on experimental situations in which a speaker uses a term that could in principle refer to two objects known to the experimental participant, but only to one object for the speaker, as the latter is unaware of the existence of the second object, and this unawareness is clear to the experimental participant. The adult participants nevertheless often perform as if the speaker referred to the object that is hidden from him, thus giving precedence to their own perspective rather than employing a first-order ToM. An example of imperfect application of second-order ToM by adults is

found in the strategic reasoning game of Hedden and Zhang (2002), which will be described in some detail in section 4.

The task dependence of successful application of ToM allows several explanations, all of which have implications for the nature of ToM. A first, and very likely, possibility is that there is a processing cost associated with ToM, which causes a failure in applying ToM or the required order of ToM when the processing demands of the task are high. Another explanation (not incompatible with the first) is that ToM does not necessarily transfer from one domain of application to another. The ability to understand another's beliefs and intentions of a certain order may be present in principle, but to apply ToM of the appropriate order, an individual must at least recognize that, in a given situation, it is to his advantage that this knowledge be incorporated in his decisions or actions. In addition, ToM may not be readily transferable from one domain to another until after a developmental process has taken place that makes this mental ability accessible to other domains, for instance Representational Redescription as proposed by Karmiloff-Smith (1992). Taking this reasoning one step further, it is even possible that what we call Theory of Mind is not one uniform mental ability to be drawn upon whenever the situation calls for it, but rather that different applications of ToM constitute different kinds of mental ability. These are all avenues of thinking about the nature of ToM that the scientific community may want to explore, however, their exploration is relevant only if first it is established to which extent there is task-dependence.

It is against this background that we place the investigations presented in this article. We compare two groups of participants, 8 to 10-year-old children and adults, on three measures. The first is a standard second-order false belief task, comparable to Tager-Flusberg and Sullivan (1994). The second is a strategic game, an adaptation of Hedden and Zhang (2002), in which participants play against a computer, trying to maximize their reward. The third measure is a linguistic task, which involves a linguistic phenomenon which is known to be acquired by children quite late, often after the age of ten. Whereas the connection between second-order ToM and the second-order false belief task will be clear, the role that second-order ToM plays in strategic games and language may not be immediately obvious. The next section will be devoted to the relation between second-order ToM on the one hand, and strategic reasoning and sentence interpretation on the other.

## 2. Theory of Mind in formal models of cognition

### 2.1 Theory of Mind and strategic reasoning in games

Games in game theory are defined by a set of players, a set of strategies available to each player, and a specification of the payoffs for each player resulting from each combination of strategies. There are two common representations for games. In normal form a game's players, strategies, and payoffs are represented in a matrix. This form is especially suitable for two-player games in which each player has only one move, and in which the players select their move simultaneously and independently. The strategies (moves) available to one player are represented as matrix rows, while the other player's strategies are represented as matrix columns. Each cell of the matrix lists the payoffs per player, if the game ends in that cell. Games may be characterized by their matrix size: A 2 by 2 game would be a game where each player chooses between two possible moves. In extensive form a game is represented as a tree, with each node representing a possible state of the game. The game starts at the initial node. Each node 'belongs' to a certain player, who chooses between the possible moves at that node. The game ends when a terminal node has been reached and the players receive the payoff specified at that terminal node. Extensive form is useful for games where players make sequential moves. Sequential games are games of perfect information: The player has complete knowledge about the actions of the other players before making his own move.

A certain game outcome (or solution) is a *Nash equilibrium* if no player can increase his payoff by choosing a different strategy while the other players keep their strategy unchanged. All finite games have at least one Nash equilibrium (Nash, 1951). Nash equilibria are easy to identify in normal form representations by looking at each player's payoffs: A cell is a Nash equilibrium if the 'column' player has no higher payoff elsewhere in the same column, while the 'row' player has no higher payoff elsewhere in the same row.

A player plays a *dominating strategy* if the strategy is better than any other strategy available, regardless of which strategy the opponent chooses. If a dominating strategy exists for a player, this strategy can be found merely by looking at that player's own payoffs without regard for the opponent's. On the other hand, a player

plays a *dominated strategy* if it is always better for him to play another strategy, regardless of what his opponent may do. If one player has a dominating strategy then all others are dominated, but the converse does not hold (see Binmore, 1992, or Osborne, 2003, for in-depth discussions of game theory).

Games can be designed so that they require particular orders of ToM for optimal performance. The use of games for ToM research has a number of advantages. First, games are different from a false belief story task in that they do not depend on language skills very much. Games are interesting because they are applied tasks. Using ToM gives the participant some advantage in the game, but the participant is not explicitly asked to use ToM, which is relevant because Keysar, Linn, and Barr (2003) showed that performance on an applied task can be far from perfect. Finally, games allow for more diversity and repetition than story tasks. As a result more items can be administered and more variation in performance between individuals can be measured.

Perner (1979) investigated children's strategies in a 2 x 2 matrix game. Although the article does not explicitly discuss ToM or order of reasoning, it can be analysed as a ToM game. The presentation of the game looked like the normal form of the game: A large wooden board was divided into four cells (two by two) with each cell containing payoffs for each of two players. The child and the opponent (an adult researcher) secretly and independently picked a row or column. After they revealed their choices the intersection of the selected row and column determined the payoff for both players. The game was designed in such a way that a dominating strategy existed for one player (the 'column player'). This player could find his optimal strategy without needing to consider his opponent's actions, so without ToM reasoning. The 'row'-player on the other hand had no dominating strategy, and could only find his optimal strategy by predicting what 'column' would do. The experiment was designed in such a way that presence of first-order ToM-reasoning could be measured.

All children played both as column and as row, and half of the children were asked to predict the opponent's choice before choosing their strategy while the other half were asked to predict after choosing their strategy. Perner found that children were more successful at picking their own dominating strategy (if the child was playing column) than at predicting that their opponent would choose his dominating strategy. The game required both first-order reasoning (when asking the child what

'column' would do) and second-order reasoning (when asking what 'row' would do). In the youngest group of 4-6 year old children only about 50% of all predictions were correct, which is consistent with chance performance. When the children's actions and predictions are crossed there are four possible outcomes. Older children were able to make correct predictions: When playing as row about 74% of all predictions were correct. However, when playing as column their performance was close to 50%. Perner thinks the children were not interested in their opponent's perspective because it did not help them: As 'column' player they had a dominating strategy that could be found without the need for prediction. However, when predicting as 'column' second-order reasoning was required rather than first-order. Thus, we propose that difficulties with second-order reasoning may also have contributed to the lower score.

An experiment designed to distinguish first- and second-order reasoning was developed by Hedden and Zhang (Hedden and Zhang, 2002). Hedden and Zhang found that adults start their game using first-order reasoning and gradually adopt a second-order strategy, but only when necessary (i.e. if their opponent is using first-order reasoning). The game was not tested on children. The application of ToM in this game may not be completely spontaneous, because participants are asked to predict the opponent's action before making their own move. Still, the results at the end of the game were far from perfect: The proportion of second-order predictions at the end of the experiment was 0.7 in the first experiment and 0.6 in the second experiment. A more in-depth analysis of the Hedden and Zhang experiment will be given in section 4.

A similar game, the so-called 'centipede game', has been studied by McKelvey and Palfrey (1992). In that game, experimental results with adults did not conform to the unique Nash equilibrium that could be computed by backward induction or the elimination of dominated strategies. Only 37 of 662 games ended with the predicted Nash outcome, in which the first player immediately moves to a dead end, causing the game to stop after only one step. Although this strategy is non-dominated, it nevertheless has a very low pay-off for the winner. In the experiments, both players would often play more cooperatively, thereby earning larger pay-offs on both sides. McKelvey and Palfrey do not explicitly use the concept of ToM reasoning in their explanation, but instead use the concepts of altruistic and egoistic reputations and incomplete information: Players may believe that there is some possibility that their opponent has payoffs different from the ones that the experimenter tries to induce by the design of the game. Although we acknowledge that concepts like

egoism and altruism can be useful in explaining participants' behaviour in strategic games, in this paper we will try to relate strategies directly to orders of ToM.

### 2.2 Theory of Mind and bidirectional Optimality Theory

In the domain of language, several phenomena have been argued to require that hearers reason about the speaker's alternatives. These phenomena include scalar implicatures, contrastive stress, object pronouns (Hendriks and Spenader, 2005/2006), and indefinite subjects and objects (de Hoop and Krämer, 2005/2006). Consider the following example of scalar implicature from Papafragou and Tantalou (2004):

(1) A: Do you like California wines?
B: I like some of them.

In this example, the term *some*, which literally means 'at least one', conveys the pragmatic interpretation of 'at least one but not all'. Although B does not literally say so, from B's answer A can conclude that B does not like all California wines. This is because the terms *some* and *all* can be placed on a scale of informativeness, with *all* being more informative than *some*. Because B did not choose the more informative term *all* on the scale, A can conclude that apparently B is not in a position to claim that the stronger form *all* is the case (for example, because using *all* would yield a false proposition). Therefore, the scalar implicature arising from B's utterance is that B does not like all California wines.

This pragmatic inference, attributed to Grice's Maxim of Quantity (Grice, 1975), has been formalized in the framework of bidirectional Optimality Theory (Blutner, 2000). According to bidirectional Optimality Theory, speakers do not merely select the best form for conveying a particular meaning, and hearers do not merely select the best interpretation for a given form. Rather, speakers also take into account the hearer's perspective, and hearers also take into account the speaker's perspective. Blutner suggests two alternative ways to account for these speaker-hearer dependencies: by means of a non-recursive mechanism of bidirectional optimization (strong optimality) or a recursive mechanism of bidirectional optimization (weak optimality). Weak optimality is defined as follows (adapted from Blutner, 2000):

(2) A form-meaning pair $<f_1,m_1>$ is bidirectionally optimal iff:

a. there is no other bidirectionally optimal pair $<f_2,m_1>$ such that $<f_2,m_1>$ is more harmonic than $<f_1,m_1>$.

b. there is no other bidirectionally optimal pair $<f_1,m_2>$ such that $<f_1,m_2>$ is more harmonic than $<f_1,m_1>$.

Bidirectionally optimal pairs are pairs for which there is no other bidirectionally optimal pair with either a better form or a better meaning. Obviously, the pair consisting of the best form and the best meaning is bidirectionally optimal. In addition, other pairs can be bidirectionally optimal as well if their competitors with either a better form or a better meaning are blocked by a bidirectionally optimal pair. Only bidirectionally optimal pairs are realized in language.

This procedure of bidirectional optimization parallels second-order belief attribution, i.e., it implies second-order ToM. When interpreting a sentence, hearers determine which meaning *m* is the best meaning for a given form $f_1$. This merely involves zeroth-order ToM. In addition, however, hearers must consider whether form $f_1$ and the selected meaning $m_1$ form a bidirectionally optimal pair, or whether an alternative form $f_2$ may express meaning $m_1$ better. Since deciding on the optimal form to express the hypothesized meaning $m_1$ requires that the hearer adopt the perspective of the speaker, this step requires first-order ToM. The hearer's belief can be represented as a first-order belief attribution, for example: "The speaker believes that meaning $m_1$ can best be expressed by using form $f_2$". If meaning $m_1$ is identified as part of a bidirectionally optimal pair $<f_2, m_1>$ containing another form than the form that was encountered, meaning $m_1$ is blocked as a possible meaning for the encountered form $f_1$. As a consequence, under weak bidirectional optimization the hearer must select a different meaning $m_2$ for the encountered form $f_1$.[1] However, this different meaning $m_2$ cannot be just any other meaning, but must be the meaning that the hearer knows the speaker believes the hearer is aware of. This can be represented as the hearer's second-order attribution about the speaker's belief: "The speaker believes that the hearer believes that alternative meaning $m_2$ is the best meaning for

[1] Under non-recursive strong bidirectional optimization, in contrast, the hearer does not have to select a different meaning for form $f_1$. Either some other meaning is equally optimal for form $f_1$ (i.e., form $f_1$ is in principle ambiguous between meaning $m_1$ and some other meaning $m_2$, such that if $m_1$ is blocked, $m_2$ remains as a possible meaning for form $f_2$), or else all pairs containing form $f_1$ are blocked. Thus, first-order ToM seems to be sufficient for applying strong bidirectional optimization.

the encountered form $f_1$ ". Thus, interpreting certain linguistic forms requires that hearers consider the alternative forms a speaker could have used, together with their associated meanings. Because the hearer must take into account the speaker's options, which in turn depend on the speaker's beliefs about what the hearer is aware of, this type of pragmatic reasoning can be argued to require second-order ToM.

De Hoop and Krämer (2005/2006) argue that errors in children's interpretation of the Dutch sentence in (3) (Termeer, 2002) are due to children's inability to optimize bidirectionally.

(3) Er ging twee keer een meisje van de glijbaan af.
there went two time a girl of the slide down
"Twice a girl went down the slide."

In De Hoop and Krämer's analysis, weak bidirectional optimization as defined in (2) accounts for adults' interpretation of the indefinite subject *een meisje* 'a girl'. The canonical word order in Dutch is one in which the subject appears in initial position. Sentence (3), however, is an existential sentence with the subject appearing sentence-internally. De Hoop and Krämer (2005/2006) argue that there is a universal linguistic constraint stating that indefinite subjects are interpreted referentially. Under a referential reading of the subject, the noun phrase *een meisje* 'a girl' refers to a particular girl rather than to any girl. Because the canonical word order (the best, or unmarked, form) expresses the referential reading (the best, unmarked, meaning), this referential reading is blocked for the marked word order in (3). As a result of bidirectional optimization, the marked word order in (3) receives a marked interpretation: a non-referential reading (any girl). To arrive at the correct reading of (3), the hearer must reason that if the speaker had wanted to express a referential meaning, he would have used the canonical word order. Upon hearing the marked word order, the hearer may conclude that apparently it was not the speaker's intention to express a referential meaning, and assign a non-referential reading to the subject *een meisje* 'a girl'. Thus, weak bidirectional optimization is crucial for obtaining the adult interpretation of (3).

Because children are not yet capable of optimizing bidirectionally, as de Hoop and Krämer (2005/2006) argue, children assign a non-adult meaning to sentence (3) and interpret the indefinite subject *een meisje* 'a girl' referentially. De Hoop and

Krämer support their explanation of children's errors in comprehension by providing developmental, language-internal and typological evidence. Crucially, because children continue to make comprehension errors with marked word order even after the age of 10, whereas they do not exhibit any problems with the comprehension of unmarked word order nor with the production of unmarked or marked word order, children's pattern of acquisition cannot be explained simply on the basis of learned form-meaning pairs.

The choice of linguistic task in our experiment is motivated by the two analyses described above. The task thus builds on the assumptions that weak bidirectional optimization requires second-order ToM, and that children's pattern of acquisition of existential sentences with an indefinite subject, such as (3), arises from the lack of weak bidirectional optimization.

Summarizing, in this section we discussed two cognitive domains where second-order ToM appears to be crucial for adult performance: strategic reasoning and language. Dekker and van Rooij (2000) show that bidirectional optimization in language corresponds to a two-player game in game theory, and bidirectional optimality to a Nash equilibrium in game theory. Thus we have a nice parallel between strategic reasoning and pragmatic reasoning: Both can be described as a sequential game between two players, and both require second-order ToM. Since children do not start out with a full-fledged ToM, the central question of this study is: How does second-order ToM reasoning develop, and how is it applied to strategic games and sentence comprehension? We will approach this question by testing how the same group of children applies second-order ToM in three different tasks: a second-order false belief task (section 3), a strategic game (section 4), and a sentence comprehension task (section 5). Although these tasks are quite different, there are control conditions for each task that do not require second-order ToM but that call upon the same cognitive functions as the conditions requiring second-order ToM. If the children perform adult-like on the control conditions of each task, we can establish to what degree the dependence on second-order ToM increases the difficulty of each task. Both the false belief task and the game task also allow comparison of the participants' performance on first-order and second-order ToM. In section 6 we will look at possible correlations between children's performance on the three tasks. Section 7, finally, presents our conclusions.

## 3. The second-order false belief test

In this section we discuss children's and adults' performance on a standard second-order false belief task.

### 3.1 Method and design

*3.1.1 Participants*

We tested 40 children from two Dutch primary schools (19 boys, 21 girls; age 8;4-10;3, mean age 9;2) and 27 adult participants (10 male, 17 female; age 18-26, mean age 20). The adult participants were psychology students participating for course credits. Each participant took part in three tests in the following order: the strategic game (discussed in section 4), the sentence comprehension task (discussed in section 5), and the false belief test (discussed below). The three tests were administered in one session that took about 30 minutes.

*3.1.2 Materials*

For the false belief test, the participants heard two second-order false belief stories, accompanied by drawings by the hand of the first author. The first story was the 'Birthday Puppy Story' reported in Tager-Flusberg and Sullivan (1994), a standard second-order false belief task. The second story, the 'Chocolate Bar Story', was a second-order adaptation of a first-order story by Hogrefe and Wimmer (1986). After each story, the participants answered several questions, modelled after Tager-Flusberg and Sullivan. The questions tested different aspects of the participant's understanding of the story, among which the participant's ability to correctly ascribe a second-order false belief such as "Mary believes that John believes that the chocolate is in the drawer". For the child group, the order of the two stories in the false belief test was balanced. The adult participants all received the Birthday Puppy Story first.

In the Chocolate Bar Story, John and Mary are in the living room when their mother returns home with a chocolate bar that she bought. Mother gives the chocolate to John, who puts it into the drawer. After John has left the room, Mary hides the chocolate in the toy chest. But John accidentally sees Mary putting the chocolate into

the toy chest. Crucially, Mary does not see John. When John returns to the living room, he wants to get his chocolate. Questions asked to the participants are: Where is the chocolate now? (reality control question), Does John know that Mary has hidden the chocolate in the toy chest? (first-order ignorance question), Does Mary know that John saw her hide the chocolate? (linguistic control question), Where does Mary think that John will look for the chocolate? (second-order false belief question), and Why does she think that? (justification question). See Flobbe (2006) for the complete texts and sets of drawings for both stories.

If the children are not able to correctly attribute second-order false beliefs but otherwise are linguistically competent, they are predicted to answer the reality control question, the first-order ignorance question and the linguistic control question correctly, but give incorrect responses to the second-order false belief question and the justification question.

### 3.2 Results

One adult gave an incorrect answer to the reality control question for the Birthday Puppy Story. The first-order ignorance question for this story was answered incorrectly by four children; the reality and linguistic control questions were answered correctly by all children. For the Chocolate Bar Story, the first-order ignorance question was answered incorrectly by one child (who also answered this question incorrectly for the other story), the reality control question was answered incorrectly by one child, and the linguistic control question was answered incorrectly by two children. All participants with an incorrect answer to any of these three types of questions (the reality control question, the first-order ignorance question, and the linguistic control question) were excluded from further analysis for that story in the second-order false belief task. The results of the remaining children and adults on both second-order false belief stories are given in Table 1 below:

| | Birthday Puppy Story | | | Chocolate Bar Story | | |
|---|---|---|---|---|---|---|
| | **N** | **Second order correct** | **Justification correct** | **N** | **Second order correct** | **Justification correct** |
| Children | 36 | 72% (26) | 56% (20) | 36 | 92% (33) | 83% (30) |
| Adults | 26 | 100% (26) | 100% (26) | 27 | 100% (27) | 100% (27) |

*Table 1: Correct responses to the second-order false belief question and the justification question for each story.*

### 3.3 Discussion

Most children responded correctly to the second-order false belief question. For the Chocolate Bar Story the correct answer to the question "Where does Mary think that John will look for the chocolate?" was "In the drawer". Many of the children also gave a correct justification for this answer, e.g., "Because Mary doesn't know that he saw that she hid the chocolate" (second-order). Children's performance on the Birthday Puppy Story is consistent with performance in the same age group found by Perner and Wimmer in their verbal second-order false belief task (1985). Children's performance on the Chocolate Bar Story is somewhat better than their performance on the Birthday Puppy Story: There was no significant difference between adults and children on the Chocolate Bar story ($\chi^2 = 2.36$, $p = 0.12$), whereas there was on the Birthday Puppy story ($\chi^2 = 8.61$, $p < 0.01$). We speculate that perhaps the Birthday Puppy Story is more difficult for children because it features more dialogue, which is not visible in the pictures. Hence the Birthday Puppy Story may tax children's memory more than the Chocolate Bar Story.

## 4. The strategic game

In this section we discuss children's and adults' performance on an adaptation of Hedden and Zhang's (2002) strategic game. Hedden and Zhang studied strategic reasoning in adults only, and used a 2-by-2 matrix game with numbers (1, 2, 3 and 4) as payoffs. Players played against an opponent in a sequential game, where first one player made a move in the matrix, and then the other player. Players were told to maximize their own payoff, and to end in a square in the matrix with the highest possible number. This required them to reason about their opponent's moves in the game. Hedden and Zhang's matrix game is, as far as we know, the only applied task that has been particularly designed to distinguish first- and second-order ToM. Because we wanted to use the strategic game to test children on their application of ToM, we had to both simplify Hedden and Zhang's game design and make it more appealing. Also, we made several improvements on their design which allowed us to rule out inappropriate transfer of simple heuristics from the training phase to the testing phase. The same participants as in the second-order false belief test participated.

### 4.1 Method and design

#### *4.1.1 Game design*

The strategic game was played on a laptop computer with a separate mouse. The participant played against a computer opponent.[2] The participant was told that he and the computer opponent were to jointly control a car. The current position in the game was represented by the location of the car. Decision points in the game were represented by road junctions. End points of the game were represented by dead ends. Each dead end contained a reward for the human player (a number of blue marbles) as well as for the computer opponent (a number of yellow marbles). The reward at a dead end could be different for each player, and the rewards to be amassed at each

[2] In Hedden and Zhang's (2002) design, one group of participants knew that they were playing against a computer, but another group was made to believe that they were playing against another participant. Hedden and Zhang found no difference in performance between the two groups. We anticipated that the deception needed in the dyad design would be extremely difficult to organize with children in a school environment, in part because all the participants are in contact with each other. Since Hedden and Zhang found that it made no difference, we chose to tell all participants the truth about their opponent.

dead end differed. Crucially, all rewards were visible throughout the entire round of the game (car ride). The reward consisted of 1, 2, 4, or 7 marbles. These numbers were chosen to make the payoffs easy to distinguish visually and to eliminate the need for counting. At each junction, the human player and the computer opponent could alternately decide either to turn to a dead end, where both drivers would receive their rewards, or to continue on the main road, so that other rewards at subsequent dead ends could be reached. Each junction was marked with a colour (blue for the human player, yellow for the computer) to show which player could decide at that junction. The participant was told to maximize his own reward (i.e., the number of blue marbles), and was told that the opponent would try to do the same (i.e., maximize the number of yellow marbles). On the left hand side of the screen, a tube gradually became filled with marbles as the human player assembled his rewards. A number, representing the score, was also displayed. There were two phases to the game. For Phase 1, first-order ToM sufficed for the participant to maximize his reward. Figure 1 shows a screenshot of the game in this phase.

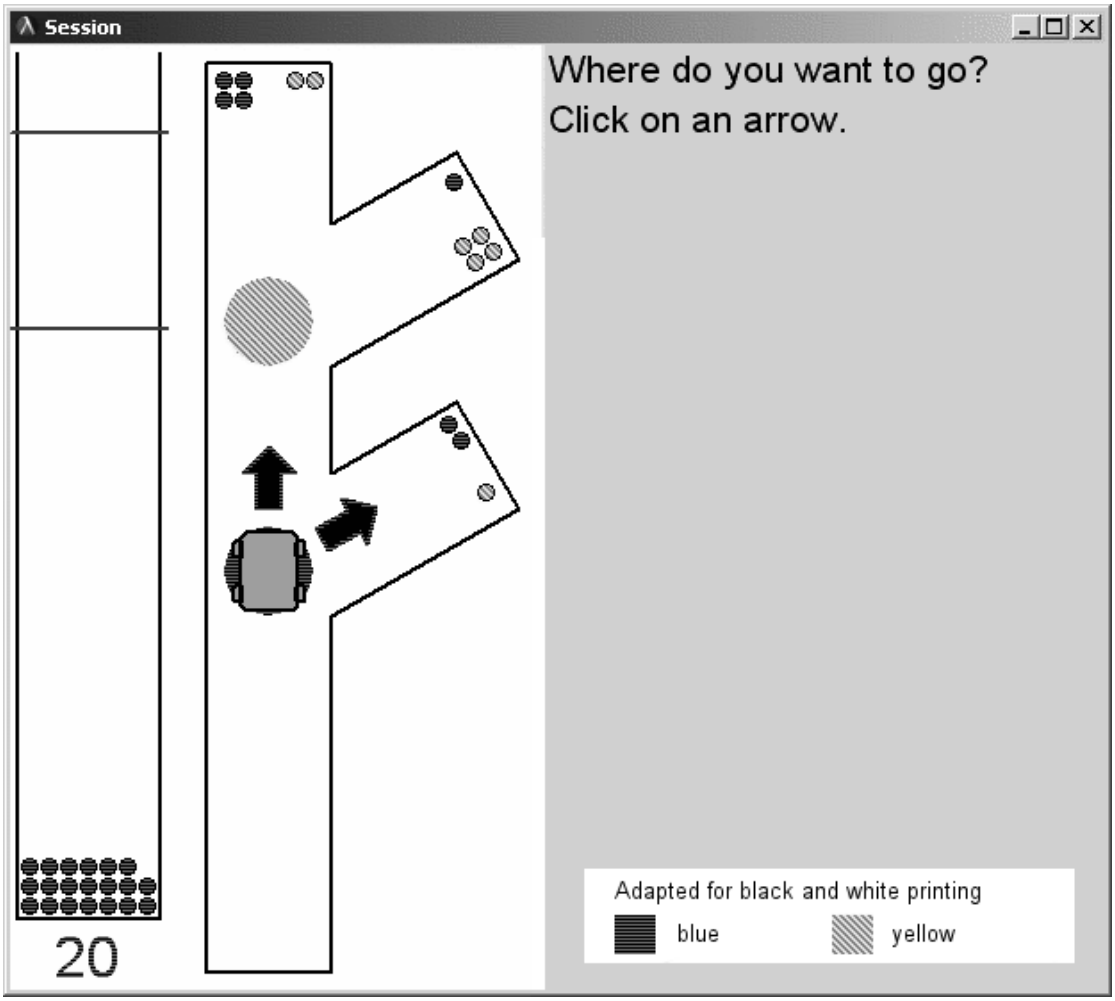


*Figure 1: A screenshot of Phase 1 of the computer program which was developed for the strategic game experiment. The human player (blue) is about to decide on his action. The tube on the left represents the human player's score.*

The human player is always the first to make a move. At the beginning of each game, the car moves to the first junction. At this point, the human player is first asked to predict the opponent's action by clicking on one of the two yellow arrows placed at the second (yellow) junction. After this the human player is asked to choose his own action by clicking on one of the two blue arrows placed at the first (blue) junction. Depending on the action chosen, the car moves ahead to the next junction or turns right to the first dead end. If the car moves to the next junction, a text message appears which indicates which action the computer opponent chooses, and the action is executed when the player acknowledges this message. When the car moves to either of the dead ends, the human player and his computer opponent will receive the reward

that is visible. The human player's reward is added to his score. A text message indicates how many marbles each player received. The message must be acknowledged before the next game is started. All car movements as well as the collection of a reward are accompanied by sounds, to increase the attractiveness of the game.

*4.1.2 Materials Phase 1*

All participants started with Phase 1, consisting of 20 items. The items in this phase had only two decision points (the first one for the human player and the second one for the computer opponent, see Figure 1) and three end points. The payoffs used were 1, 2, and 4, which were distributed over the three end points. The first 4 items of Phase 1 were familiarization items, in which the participant was not asked to make predictions. In the remaining 16 items, the participant had to first predict the opponent's next action before making his own move. The items included all 12 different combinations in which the human player started with a payoff of 2 at the first dead end, and 8 games in which the human player started with a payoff of 1 or 4 at the first dead end (see Flobbe, 2006, for a complete list of items).

Phase 1 served two purposes: It functioned as a training session, and also allowed us to determine whether the participants were capable of first-order ToM reasoning. Consider the situation depicted in Figure 1. If the participant is capable of first-order ToM reasoning, he will be able to correctly predict the opponent's action in the second move. Assuming that the opponent also tries to maximize his reward, having arrived at the yellow junction, the opponent will turn right to the second dead end, which yields 4 yellow marbles, rather than move straight ahead to the third dead end, which would yield only 2 yellow marbles. A participant capable of applying first-order ToM in strategic reasoning will be able to use this prediction to rationally determine his own action. In the situation depicted in Figure 1, the best action for a human player at the blue junction would be to turn right to the first dead end (which yields 2 blue marbles) rather than move straight ahead to the yellow junction. The latter move would yield only 1 blue marble, given the first-order ToM prediction that the opponent will turn right at the yellow junction.

A first-order strategy requires players to take into account their opponent's desires. It assumes that the opponent acts as a zeroth-order player, who only takes into account his own desires and the state of the world and simply chooses the largest

payoff at that position. If a human player does not apply first-order ToM reasoning, several zeroth-order strategies are possible: averaging over the rewards, heading for the maximal reward, or simply random behaviour. Crucially, for the items in Phase 1, second-order reasoning is not useful and would lead to the same result as first-order reasoning.

The last 6 items of Phase 1 were treated as test items for our analysis: They allowed us to determine whether children applied first-order ToM reasoning. We assumed that all children who participated in the experiment would be capable of first-order ToM on a standard task. Most children pass the first-order false belief test at age 4 according to Wimmer and Perner (1983), and the first-order components of the standardized Dutch ToM-test (Steerneman, Meesters, and Muris, 2003) have a success rate of over 70% by age 8. Whether our child participants would also be able to apply first-order reasoning in the game task was one of the questions this study had set out to answer. Determining whether participants were capable of applying first-order ToM reasoning in Phase 1 was also essential for interpreting the results of Phase 2, which required second-order reasoning. Since the items in Phase 2 were designed to distinguish first-order reasoning from second-order reasoning, participants who are not even capable of first-order reasoning should be excluded from analysis.

#### *4.1.3 Materials Phase 2*

Phase 2 consisted of 4 sets of 10 items each. Of these 40 items, 32 were diagnostic items, which allowed us to distinguish first- and second-order strategies. The remaining 8 items were control items, for which a first- and second-order strategy would yield the same predictions. The items in this phase had three decision points - the first one for the human player, the second one for the computer opponent, and the third one for the human player again (see Figure 2) - and four end points. The payoffs were 1, 2, 4, and 7. Preceding the 40 items of this session, participants started with 4 items for familiarization, in which the participant was not asked to predict the move of the opponent. The test items consisted of combinations in which the human player started with a payoff of 2 or a payoff of 4 at the first dead end (cf. Hedden and Zhang, 2002) in a random order (see Flobbe, 2006, for a complete list of items).

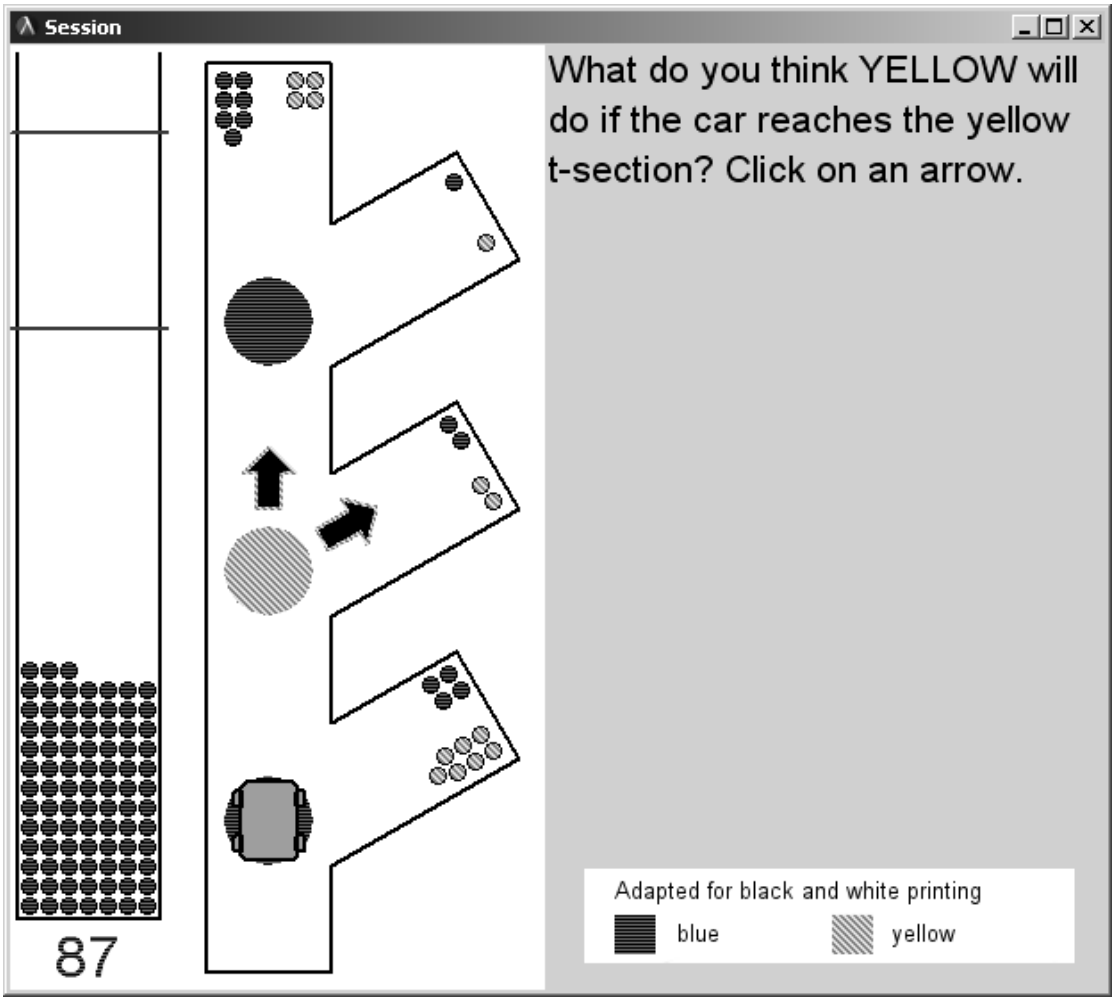


*Figure 2: Another screenshot of the computer program developed for the strategic game experiment. This screenshot shows Phase 2.*

In Phase 2, the computer opponent always used first-order reasoning (in contrast to Hedden and Zhang's experiment, where participants played either against a zeroth-order 'myopic' player or a first-order 'predictive' player). Consider the situation in Figure 2. First, the human player is asked to predict the action of the computer opponent at the yellow junction. If the human player uses a second-order strategy, he will assume that the computer opponent acts as a first-order player. This first-order opponent will assume that the human player acts as a zeroth-order player at the last (blue) junction, and that he will move straight ahead to the fourth dead end to receive the large reward of 7 blue marbles. The second-order human player will predict that the first-order opponent will move straight ahead at the second yellow junction, as the fourth dead end, which can be reached from the next junction, not only contains the

largest reward for the human player, but also for the computer opponent. The second-order human player knows that his first-order opponent is aware of his (the human player's) desire to gain the largest reward, and that the first-order opponent will count on this desire in passing the turn to the human player again, rather than turning right to gain a mere 2 yellow marbles. At the first junction, a second-order human player has to compare the reward at the fourth dead end (7 blue marbles) with the reward at the first dead end, which he will receive if he decides to turn right (4 blue marbles). A second-order human player will therefore always decide to move straight ahead at the first junction, counting on the first-order opponent allowing him to "turn into the street" that has the largest rewards for both of them. In game-theoretic terms, the second-order player uses backward induction to eliminate dominated strategies, thereby attaining the Nash equilibrium, similarly as in the rational solution to the centipede game (cf. McKelvey & Palfrey, 1992).

If, on the other hand, the human player uses a first-order strategy, he will assume that the computer opponent acts as a zeroth-order player at the second (yellow) junction. A zeroth-order player will not take into account the opponent's desires but only act upon his own desires. As a result, a zeroth-order opponent may decide to turn right or move straight ahead, depending on the exact heuristic employed. Because the reward for the human player at the second dead end on the right (2 blue marbles) is smaller than the reward at the first dead end (4 blue marbles), a human player using a first-order strategy may therefore decide to turn right at the first junction.

### 4.2 Results

The last 6 items of Phase 1 were used to determine whether participants were capable of at least first-order ToM reasoning in this task. If a participant made an incorrect prediction, this was counted as a prediction error. If the participant made a correct prediction but nevertheless chose an action that did not maximize his payoff, this was counted as a rationality error. Figure 3 shows the proportion of errors that were made during the last 6 training items.

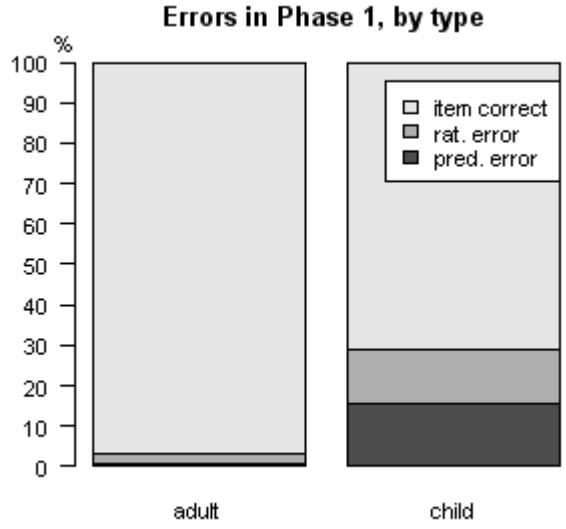


*Figure 3: Proportion of prediction errors, rationality errors, and correct actions for the last 6 items of Phase 1, testing first-order reasoning.*

Adults chose the correct action for 97% of the final 6 items of Phase 1, whereas the children chose the correct action for 71% of the items. Of the 27 adults, only 1 made more than one mistake. Of the 40 children, 18 made more than one mistake. Nine of these 18 children made no more than one prediction error.

All participants who made more than one error in the last 6 items of Phase 1, 1 adult and 18 children, were excluded from the analysis of the results of Phase 2. One additional child was excluded when we took a closer look at the player's decision at the last junction in items from Phase 2. Here, the player does not need to reason about his opponent's actions anymore and simply has to choose the highest reward of the two. However, several children did not choose the highest reward at this point.[3] We decided to exclude players who selected an incorrect action at the last junction in more than 20% of the cases, with a minimum of three incorrect actions. Consequently, 10 children (9 of which were children that were already excluded by the previous

[3] Interestingly, further analysis of these cases revealed that in all of these cases the incorrect action chosen by the child maximized the difference between the payoff for the player and the payoff for the computer opponent. This suggests that these children tried to collect more marbles than their opponent, hence entertaining inappropriate competitive goals. We had considered this possibility after a pilot phase with four adult participants. To avoid competitive behaviour, we had made a number of changes to the design of the game: We displayed the participant's score prominently, while hiding the opponent's score, and we also displayed two target scores (the two horizontal lines at the top of the tube in Figure 1 and 2), which yielded a real reward (a sticker for children, candy for adults) when reached. Furthermore, we emphasized in our instructions to the participants that the opponent's score did not influence their rewards. However, these revisions may not have been sufficient to completely prevent all children from entertaining inappropriate competitive goals.

criterion of correct first-order reasoning) were excluded from further analysis. This leaves us with 26 adults and 21 children in the analysis of Phase 2. Figure 4 shows the proportion and type of errors that were made in Phase 2.

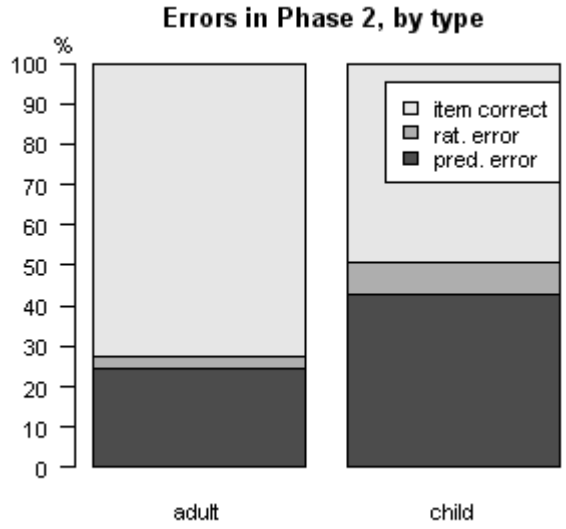


*Figure 4: Proportion of prediction errors, rationality errors, and correct actions for all items in Phase 2, testing correct second-order reasoning.*

Phase 2 consisted of 32 diagnostic items, presented in four sets. For each of the remaining participants, the number of correctly predicted items was calculated and divided by 32 to obtain the percentage of correctly predicted items. The mean correct prediction for children was 57.2% (score 18.29, SD = 5.68), and for adults 75.5% (score 24.15, SD = 5.62). The child mean is significantly higher than the mean of 16 that we would expect if all participants were guessing (one sample t-test, $t = 1.85$, one-sided $p = 0.04$). In Figure 5, the individual prediction scores for each participant are presented in a histogram.

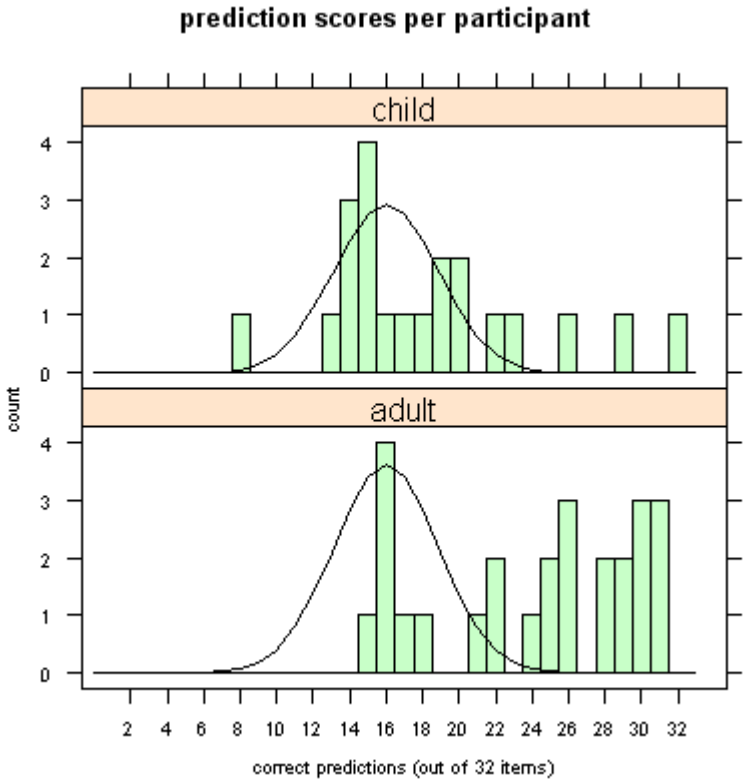


*Figure 5: Histograms showing the prediction scores for each participant in Phase 2, testing correct second-order reasoning. The maximum obtainable prediction score was 32. The black curve represents the (binomial) distribution of scores that is expected if participants are guessing.*

A cursory look at the data gives the impression that some individuals, especially in the child group, score around chance level.[4] It must be noted, however, that it is unlikely that participants were guessing, since participants who had not demonstrated the ability to correctly apply first-order reasoning in Phase 1 were excluded from analysis. However, section 4.3 explains that there are other possible explanations, besides guessing, for a score around 50%.

Since the test items were presented in four subsequent sets of 8 items, the prediction score per set could be used to determine whether performance changed during the experiment. The adults showed a rather small but statistically significant ($\beta$

[4] It is not possible to divide the population into those who score at chance level and those who score above chance level. The hypothesis that a particular individual score derives from chance can be rejected for those individuals who have answered 22 or more of the 32 items correctly: $p(x \geq 22) < 0.025$ while $p(x \geq 21) < 0.055$, calculated from the binomial distribution B(32; 0.5). This is the case for 5 children and 18 adults. However, it would be a fallacy to conclude that all other participants score at chance level.

= 3.5%, $p = 0.0095$) increase in correct prediction rate during the experiment. The children showed a small decrease during the experiment, but this was not significant ($\beta = -3.0\%$, $p = 0.056$).

Most of the time participants chose an action that was consistent with their prediction. However, sometimes an incorrect action was chosen despite a correct prediction. These rationality errors constitute 7.7% of all items for children and 3.1% for adults (as can be seen in Figure 4), which is 13.5% and 4.1%, respectively, as a proportion of correctly predicted items.

### 4.3 Discussion

From the results of Phase 1 we can conclude that the majority (77%) of the tested children are capable of making first-order predictions, although these predictions are not always used to select the correct action. In Phase 2 we found that children perform above chance with second-order ToM reasoning, with a success rate of 57.2%, but clearly below the adult group. It should be kept in mind that those participants with low success rates on first-order ToM were excluded from further analysis based on their scores on Phase 1. If we compare the scores on first- and second-order ToM items for only those children who were included in Phase 2, the difference is even more striking: These child participants had 57.2% correct predictions on second-order ToM items, compared to 93% correct predictions on the first-order ToM items. Clearly, children find second-order ToM more difficult than first-order ToM in the game task. However, the adult success rate on the second-order ToM items (75.5%) shows that even adults do not reliably apply second-order ToM reasoning when needed.

In Hedden and Zhang's experiment (2002), adult prediction scores started at a low value of around 20% for the first item set, and then rose to about 60-70% towards the end of the test session. Our experiment yields different results: The adults have a prediction score around 75% throughout Phase 2, with only a small increase in performance during the experiment. In other words: Our adult participants perform better than Hedden and Zhang's participants, and they do not improve much during the experiment. We offer two possible explanations for this difference. The first explanation is that our experiment uses a different and more concrete presentation.

These changes were made so that the game could be played by children, but they may have helped adults as well. Since the test items are mathematically equivalent to those of Hedden and Zhang, the difference should not be important once a participant is thoroughly familiar with the game. But especially at the start of the game, a better presentation may improve performance. A second explanation is that the difference in results between our Phase 2 and Hedden and Zhang's test session is caused by a difference between the items of our Phase 1 and Hedden and Zhang's training items. Hedden and Zhang use a special class of 'easy' items with 4 cells (dead ends in our presentation) in their training session, for which first- and second-order predictions should give the same results. We believe that these items may have encouraged participants to use the 'easier' first-order strategy and later try to apply this incorrect strategy to the superficially very similar testing items. If we are correct, the improvement of Hedden and Zhang's results during the test session represents 'unlearning' an inappropriate strategy. Our items in Phase 1 are quite different from Hedden and Zhang's training items, as they have only 3 dead ends, which makes them visually distinct from the items in Phase 2. Participants will immediately notice that the items in Phase 2 are more complicated than the items in Phase 1, and that the strategy used during Phase 1 cannot be applied to Phase 2. The present findings call into question Hedden and Zhang's conclusion that adults use first-order ToM as a default, only moving to second-order when the need arises, thus initially crediting their opponents with no more than zeroth-order ToM.

The prediction scores of our participants in Phase 2 should be interpreted with care because a first-order player may entertain different assumptions about how a zeroth-order opponent would act, giving rise to different predictions. Colman (2003) pointed out that Hedden and Zhang's characterization of zeroth-order ('myopic') behaviour by the computer opponent is problematic. A zeroth-order player only takes into account his own payoffs while disregarding his opponent's payoffs and options. Hedden and Zhang's myopic opponent compares his payoffs at the second and third cell to decide where to move at the second junction. Colman points out that there are various ways in which a zeroth-order opponent could take into account his payoff at the fourth cell. He could average the third and fourth cell payoffs, or he could look at the maximum or the minimum in the third and fourth cells. Although we did not implement a zeroth-order opponent in our own experiment, the critique voiced by Colman is still relevant. A goal of Phase 2 was to distinguish second-order reasoning

from first-order reasoning. If we accept different assumptions about how a zeroth-order opponent would act, we should also expect different responses from participants who use first-order reasoning. We examined alternative first-order strategies, based on alternative assumptions about how a zeroth-order opponent would act, and found that a first-order reasoner could have answered up to half of all items of Phase 2 correctly (i.e., like a second-order reasoner). Therefore, we cannot claim that a given prediction score by a participant represents a specific proportion of first-order and second-order reasoning. We used items from Hedden and Zhang for which second-order reasoning would lead to a different response than Hedden and Zhang's proposed first-order strategy. Since each item can only have two possible responses, it would not have been possible to accommodate other first-order strategies in the experiment as well. Therefore, we need to consider the possibility that the participant may use some strategy that allows him to answer up to 50% of the items correctly, without using second-order reasoning. Prediction scores significantly higher than 50%, however, are indicative of second-order reasoning. In our experiment, both the child and adult group had a mean prediction score significantly higher than 50%, which indicates that both groups used second-order reasoning to at least some degree.

In general, both adults and children perform better in Phase 1 of the game, when only first-order reasoning is required, than in Phase 2, when second-order reasoning is required. Adults perform significantly better than children do. These findings are consistent with the idea that second-order reasoning develops at a later age than first-order reasoning.

## 5. The sentence comprehension test

In this section we discuss children's and adults' performance on a sentence comprehension test with indefinite subjects (cf. Termeer, 2002; Vrieling, 2006).

### 5.1 Method and design

#### *5.1.1 Participants*

The same participants as in the second-order false belief test and the strategic game participated. All children were native speakers of Dutch. Two adult participants were

excluded from the sentence comprehension test because they were not native speakers of Dutch.

*5.1.2 Materials*

The test materials were taken from Vrieling (2006). Participants heard two stories, in each of which two different girls perform a certain action. After each story the participant heard a sentence and had to decide whether this sentence was correct. We tested the comprehension of two types of sentences:

(4) Een meisje ging twee keer van de glijbaan af. (canonical sentence)
a girl went two time of the slide down
"A particular girl went down the slide twice."

(5) Er ging twee keer een meisje van de glijbaan af. (existential sentence)
there went two time a girl of the slide down
"Twice a girl went down the slide."

Each participant heard one canonical sentence and one existential sentence. The items were balanced so that half of the participants received an existential sentence first, and the other half a canonical sentence first.

De Hoop and Krämer (2005/2006) argue that sentence (5) requires the hearer to take into account the speaker's perspective and reason about alternative, unheard forms, such as sentence (4), and their meaning. The reasoning proceeds as follows. Sentence (4) is the unmarked form, because the subject appears in its canonical position. Furthermore, there is a cross-linguistic tendency for indefinite subjects as in (4) to preferably be interpreted as expressing a referential reading ('a particular girl'). Because a referential reading is the preferred reading for indefinite subjects, a referential reading should also be the preferred reading for existential sentence (5). However, a hearer can reason that if the speaker had wanted to express the unmarked referential meaning, he would have produced the unmarked, canonical sentence form in (4). Consequently, upon hearing existential sentence (5), the hearer concludes that apparently it is not the speaker's intention to express a referential meaning, and assigns a non-referential reading to the subject in (5).

If young children are incapable of this type of reasoning about the speaker's options (as is argued by de Hoop and Krämer, 2005/2006, and Hendriks and Spenader,

2005/2006), we predict differences between children and adults when comprehending (5) but not (4). In particular, we predict that children will assign an interpretation to existential sentence (5) according to which it must be the same girl who went down the slide (a referential reading). Such erroneous interpretations were indeed found by Termeer (2002) and Vrieling (2006).

### 5.2 Results

The results of the sentence comprehension task are given in Figure 6. The difference between adults and children for the existential sentences is highly significant ($\chi^2$ = 23.78, $p < 0.00001$).

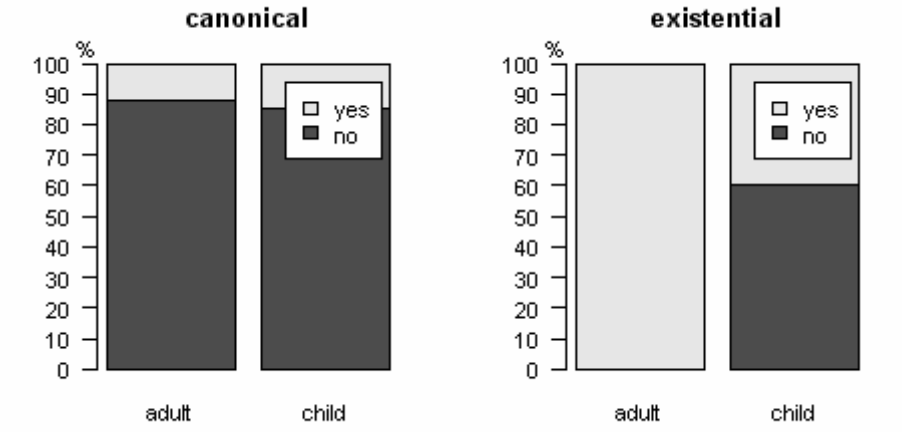


*Figure 6: Correctness judgements for canonical sentences and existential sentences. For canonical sentences, the grammatical response is "No", for existential sentences, the grammatical response is "Yes".*

### 5.3 Discussion

The adults always assign a non-referential reading to existential sentences such as (5): Two different girls may have gone down the slide. Most children (24 out of 40), in contrast, preferred a referential reading for an existential sentence: It must be the same girl who went down the slide. Canonical sentences such as (4) were interpreted identically by children and adults. Only 3 out of 25 adults and 6 out of 40 children assigned a non-referential reading to a canonical sentence. This outcome is as predicted by de Hoop and Krämer (2005/2006). If their analysis of indefinite subjects

is correct, our results indicate that most nine-year-old child hearers are not yet able to reason about the speaker's options with respect to indefinite subjects.

## 6. General discussion

In this section we discuss how performance on the three tasks is related. Because of the rather uniform, high performance on the second-order false belief task, an analysis of the relationship between performance on the second-order false belief task and on the strategic game or the sentence comprehension task cannot be statistically significant. Only three children answered the second-order false belief question about the chocolate story incorrectly, and only one of these three children was included in the analysis of Phase 2 of the strategic game. The results are consistent with the assumption that passing a second-order false belief task is a necessary condition for applying second-order reasoning to the strategic game, but this assumption cannot be proven because there is insufficient variation in the data. Similarly, passing a second-order false belief task may be necessary for applying weak bidirectional optimization in the sentence comprehension task, but because of lack of variation in the results on the false belief task we cannot draw any conclusions about this assumption either.

To investigate a possible link between the sentence comprehension task and the strategic game, we divided the children into two groups by their response to the existential sentence, and compared the average prediction scores on the strategic game for these groups. The response to the canonical sentence was not informative, because the proportion of 'deviant' responses to this sentence was very low and similar for adults and children. We did not include the adult data in our analysis. Given that all adults gave the same response to the existential sentence and adults performed better on the strategic game than children did, inclusion of adult data in the analysis might give spurious results. Of the children included in Phase 2 of the strategic game, 8 gave an adult-like 'yes'-response to the existential sentence, and 13 gave a non-adult-like 'no'-response. For the children who were excluded from Phase 2, the proportion of responses was similar. The children with an adult-like response had a mean prediction score of 50%, while the non-adult-like group had a mean prediction score of 61%. A two sample t-test with pooled variance found that the difference between these means is not significant ($t = 1.49$, $p = 0.15$). Despite sufficient variation in the data for each

task, we did not find a correlation between a child's response on the sentence comprehension task and a child's prediction score on the strategic game.

What can we conclude from the absence of a correlation between the responses on the sentence comprehension task and the score on the strategic game? Apparently, applying ToM is not a unitary skill that develops independently of the domain of application. Rather, learning to apply second-order reasoning appears to crucially depend on the domain of application, which can be a strategic game, sentence comprehension, or some other domain where taking into account other people's mental states may be useful. In particular the discrepancy between the results on the standard second-order false belief task and the strategic game task raises a host of questions, as nearly all children passed the second-order false belief task but only just over half applied second-order ToM in the strategic game. We will mention just one of these questions: What is it that children need for their performance on the strategic game task to improve? It could be that mere practice would suffice, if the strategic game takes up too many processing resources. This however raises the question of why processing and judging the situation in the standard second-order false belief task is so much easier – formally, the tasks are equal, and we might even argue that the strategic game task should be easier as a real gain is to be obtained from applying second-order ToM.

This brings us to another possibility: Children may need to learn to recognize the importance of applying second-order ToM in the situation of the game. However, this once more leads to the question of what it is about this game situation that makes this more difficult than in the situations sketched in the second-order false belief task.

One testable hypothesis that addresses both of these possibilities is that the abstraction involved in the game task is the key factor, i.e. that the child participants, and possibly also the adults, can apply second-order ToM more easily in situations that involve actual people. Their representation or physical presence could support the knowledge that the other has goals and desires, and also has insight into other people's goals and desires. This would point to the relevance of the social embedding of Theory of Mind abilities, and possibly the role of social interaction in its emergence. This, and other questions that present themselves, we leave for future research.

## 7. Conclusion

We used an adapted version of Hedden and Zhang's (2002) strategic game to test children on their application of ToM. The majority of 8-10 year old children and all adults were able to make correct first-order predictions at the end of the first phase of our version of the strategic game. After excluding participants who did not correctly apply first-order reasoning, the remaining participants demonstrated second-order reasoning in the second phase of the strategic game. Adults make more correct second-order predictions (75.5%) than children (57.2%) do. However, in both groups performance is far from perfect. Our results differ from Hedden and Zhang in that we did not find a learning effect or strategy change during the game. Participants who used second-order reasoning did so from the start of the second phase of the game. Adults perform better on the strategic game than the children. We can think of two reasons why this is so, the most likely of which is that applied ToM reasoning continues to develop after the age of 8-10 years. Another possibility is that IQ or factors related to IQ play a role – as the adults were university students, we may assume that they have above average intelligence. This is a possibility which we leave to future research.

In addition to testing children on a strategic game, we also tested children on their application of ToM on a second-order false belief task and a sentence comprehension task. Children's application of second-order ToM was found to be highly dependent on the task to be carried out and the domain of application. Whereas almost all children succeeded on a verbal second-order false belief task, children's success rate in our second-order strategic game was only 57.2%. With respect to the sentence comprehension task, only 40% gave a bidirectionally optimal interpretation of the indefinite subject of an existential sentence. Despite sufficient variation in the data for the strategic game and the sentence comprehension task, we found no relation between children's performance on the strategic game task and on the linguistic task.

Thus, we have found that second-order ToM is more difficult to apply than first-order ToM, for children as well as adults, and that this pattern not only holds for verbal false belief tasks, but also for strategic games. Moreover, we have found that successful application of second-order ToM depends crucially on the domain in which it must be applied. This finding shows that, beyond the question of how human beings

come to *have* a Theory of Mind, there looms another important question: How do we learn to *use* it?

**Acknowledgements**

We thank the children and staff of the St. Jorisschool in Heumen and the Christelijke Basisschool De Bron in Marum for their cooperation. Pauline Vrieling and Daniëlle Koks were so kind as to allow us to use their materials for our language experiment. We also thank the participants of the workshop "Formal models for real people" and two anonymous reviewers of this journal for their valuable suggestions and comments. Rineke Verbrugge gratefully acknowledges the NIAS (Netherlands Institute for Advanced Studies in the Humanities and Social Sciences) for awarding her a fellowship in the framework of the project 'Games, Action, and Social Software'. Furthermore, we gratefully acknowledge the Netherlands Organisation for Scientific Research, NWO (grants no. 051-04-120 and 400-05-710 for Verbrugge, grant no. 051-02-070 for Krämer and Hendriks, and grants no. 277-70-005 and 015-001-103 for Hendriks).

**References**

Binmore, K (1992). *Fun and Games: A Text on Game Theory*. Lexington (MA): D.C. Heath and Company.

Blutner, R. (2000). Some aspects of optimality in natural language interpretation. *Journal of Semantics*, 17, 189-216.

Colman, A.M. (2003). Depth of strategic reasoning in games. *Trends in Cognitive Sciences*, 7, 2-4.

De Hoop, H. & Krämer, I. (2005/2006). Children's optimal interpretations of indefinite subjects and objects. *Language Acquisition*, 13, 103-123.

Dekker, P. & Van Rooij, R. (2000). Bi-directional optimality theory: An application of game theory. *Journal of Semantics*, 17, 217-242.

DeVilliers, J. (2007). The interface of language and Theory of Mind. *Lingua*, 117, 1858-1878.

Flobbe, L. (2006). *Children's development of reasoning about other people's minds*. MSc Thesis Artificial Intelligence, University of Groningen.

Grice, H. P. (1975). Logic and conversation. (In: P. Cole & J.L. Morgan (Eds.), *Syntax and Semantics, vol. III, Speech Acts* (pp. 41-58). New York: Academic Press.)

Hedden, T. & Zhang, J. (2002). What do you think I think you think? Strategic reasoning in matrix games. *Cognition*, 85, 1-36.

Hendriks, P. & Spenader, J. (2005/2006). When production precedes comprehension: An optimization approach to the acquisition of pronouns. *Language Acquisition*, 13:4, 319-348.

Hogrefe, G. & Wimmer, H. (1986). Ignorance versus false belief: A developmental lag in attribution of epistemic states. *Child Development*, 57, 567.

Karmiloff-Smith, A. (1992). *Beyond Modularity: A Developmental Perspective on Cognitive Science*. Cambridge (MA): MIT Press.

Keysar, B., Lin, S., & Barr, D. (2003). Limits on theory of mind use in adults. *Cognition*, 89, 25-41.

McKelvey, R.D. & Palfrey, T.R. (1992). An experimental study of the centipede game. *Econometrica*, 60, 803-836.

Mol, L., Verbrugge, R. & Hendriks, P. (2005). Learning to reason about other people's minds. (In L. Hall, D. Heylen et al. (Eds.), *Proceedings of the Joint Symposium on Virtual Social Agents* (pp. 191-198). The Society for the Study of Artificial Intelligence and the Simulation of Behaviour (AISB), Hatfield.)

Nash, J. (1951). Non-cooperative games. *The Annals of Mathematics*, 54, 286-295.

Onishi, K. & Baillargeon, R. (2005). Do 15-month-old infants understand false beliefs? *Science*, 308, 255-257.

Osborne, M. (2003). *An Introduction to Game Theory*. Oxford: Oxford University Press.

Papafragou, A. & Tantalou, N. (2004). Children's computation of implicatures. *Language Acquisition*, 12, 71-82.

Perner, J. & Ruffman, T. (2005). Infants insight into the mind: How deep? *Science*, 308, 214-216.

Perner, J. & Wimmer, H. (1985). "John thinks that Mary thinks that …": Attribution of second-order beliefs by 5- to 10-year-old children. *Journal of Experimental Child Psychology*, 39, 437-471.

Steerneman, P., Meesters, C., & Muris, P. (2003). *TOM-test* (derde druk). Antwerpen: Garant.

Tager-Flusberg, H. & Sullivan, K. (1994). A second look at second-order belief attribution in autism. *Journal of Autism and Developmental Disorders*, 24, 577-586.

Termeer, M. (2002). *"Een meisje ging twee keer van de glijbaan." A study of indefinite subject NPs in child language*. MA Thesis, Utrecht University.

Vrieling, P. (2006). *Een ezel stoot zich geen twee keer aan dezelfde steen: Dutch children's interpretation of indefinite subject NPs*. MA Thesis, Utrecht University.

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