Children's development of reasoning about other people's minds

MSc Thesis

by
Liesbeth Flobbe
stud.nr.: xxxxxxx

Supervised by:
Petra Hendriks[1]
Irene Krämer[2]
Rineke Verbrugge[3]

Artificial Intelligence
University of Groningen

November 2006

[1]Dutch Language and Culture, University of Groningen, E-mail: p.hendriks@rug.nl
[2]Linguistics, Radboud University Nijmegen, E-mail: ikramer@let.ru.nl
[3]Artificial Intelligence, University of Groningen, E-mail: l.c.verbrugge@ai.rug.nl

Children's development of reasoning about other people's minds

by Liesbeth Flobbe

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 reasoning is the false belief task. A different approach is used by Hedden and Zhang (2002), who investigated the application of ToM reasoning in a strategic game. Another task that is thought to involve second order ToM is the comprehension of sentences that the listener can only understand by considering the speaker's alternatives.

In this research a group of 8-10 year old children and a group of adults were tested on (adaptations of) the three tasks described above. The results show interesting differences between adults and children, between the three tasks, and between this experiment and previous research.

Thesis supervisors: L. C. Verbrugge, P. Hendriks, I. Krämer

Acknowledgements

I would like to thank the children and staff of the St. Jorisschool in Heumen and the Christelijke Basisschool de Bron in Marum. Not only would this research not have been possible without them, I also greatly enjoyed the time I spent at their schools.

I would like to thank Paulien Vrieling and Danielle Koks for allowing me the use the drawings for the indefinite subject experiment.

I would like to thank my supervisors Rineke Verbrugge, Petra Hendriks, and Irene Krämer for their invaluable advice.

Finally I would like to thank John for proofreading and for all the support he has given.

Contents

Chapter 1

Introduction

Cognitive science studies human intelligence. In this thesis I study the development of a specific aspect of human intelligence: the ability to reason about the knowledge and intentions of other people. I will adopt the approach known as Theory of Mind (ToM), which understands this development by assuming that children develop a 'theory of mind': a mental model of other people's minds. While a lot of research has focused on very early development of theory of mind, this work will rather focus on the advanced application of theory of mind. One area in which people apply their theory of mind is in the area of strategic games. The hypothesis behind the current project is that theory of mind is also applied in a very different area: in the comprehension of certain sentences. There are a number of linguistic constructions for which correct comprehension develops at a relatively late age, and these constructions have in common that the listener must reason about the speaker's alternatives to understand the speaker's intended meaning. I will use the mechanism of bidirectional optimization, an expansion to the linguistic model called Optimality Theory, to describe these linguistic phenomena and to explain the resemblance with ToM-reasoning.

The research question for this project is:

> How does children's development of the ability to reason about other people's knowledge and intentions correlate with the development of the ability to reason about speaker's alternatives during language comprehension?

The main objective of this project is to prove or disprove a link between theory of mind and language. The main method to achieve this goal is experimental research, on both adults and children. If a link between theory of mind and language is found, this will strengthen the theoretical justification for bidirectional optimization and contribute to our knowledge of the language faculty. But the experiments on theory of mind will also by themselves contribute to our knowledge of advanced theory of mind and its development. It should be evident that research on theory of mind, as an aspect of human intelligence, is important for cognitive science. Insights from ToM-research are already being used

in clinical practice, in diagnosing individuals with specific impairments such as autism. Hopefully some day ToM-research will also tell us how to help these individuals. Finally, ToM-research can inform Artificial Intelligence, especially the field of multi-agent systems, which aims to develop robotic or software agents that can reason about other agents, including humans.

In chapter 2 I will give a summary of the relevant literature on theory of mind, linguistics, and game theory. Chapter 3 will then revisit the research question. I will summarize the partial answers to the research question that can be found in the literature, and will formulate more precise questions to guide the experimental part of this research. The experimental design will be described in chapter 4. The final chapters then present the results of the experiments, a discussion, the conclusion, and a summary.

Chapter 2

Theoretical background

In this thesis I investigate how Theory of Mind may be involved in pragmatic language use and in playing strategic games. In the first section I will give an overview of Theory of Mind. In the second section I will describe a number of linguistic phenomena that involve reasoning about the speaker's alternatives. In the final section I will give an introduction to game theory and describe how games have been used in ToM-research.

2.1 Theory of Mind

2.1.1 What is Theory of Mind?

Many everyday reasoning tasks require a person to reason about the knowledge and intentions of other people. The capacity for this kind of reasoning is called *mind reading*. A common approach to studying this capacity uses the phrase *theory of mind* (ToM) for it. This phrase was 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 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 distinct from his own. He can formulate hypotheses about what those beliefs, desires, and intentions are.

ToM-reasoning can be classified by its order of reasoning. Reasoning about other people's beliefs is usually first order reasoning. Examples of first order statements are: "(I know that) Mary thinks the ball is in the bag." or "(I know that) you intend to take the left cup." However, if a person takes into account the other person's beliefs about the minds of others (including his own), that person uses second order reasoning. Examples of second order statements are: "(I know that) Mary thinks that John thinks the ball is in the cupboard." or "(I know that) you think that I think the box contains a pencil."

In other words, a second order reasoner thinks of other people as first order reasoners. Somebody who thinks about others as second order reasoners would himself be exhibiting third order reasoning. Higher levels of reasoning can be constructed ad infinitum, but in most situations people cannot cope with more than second order reasoning. Second and higher order reasoning are also called recursive reasoning.

There are more distinctions within ToM-reasoning than just the order of reasoning. Most studies have investigated reasoning about beliefs, but ToM-reasoning can also be about intentions, desires, goals, or any other propositional attitude.

2.1.2 The development of ToM

Research of infant development has uncovered many 'precursors' of ToM. At 9 months infants are able to follow an adult's eye-gaze and establish 'joint attention' towards an object. Children of 2 years engage in pretend play. 3-year-olds sometimes seem to deceive others (see Flavell and Miller, 2002 for an overview). Even very young children are able to guess the intentions of another. The Rubicon of ToM is however the false belief task, first used by Wimmer and Perner (1983). In a false belief task the child is asked to predict the behaviour of another person, for example to predict 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 other people have beliefs, and that the child can distinguish between its own beliefs and those of others. Children at age 3 still fail false belief tasks, but children at age 4 or older pass them.

The idea that children under 4 years of age have no ToM at all has come under attack by a number of researchers. Although the verbal predictions that 3-year-olds make in false belief tasks are 'wrong', their eye gaze direction indicates an understanding of false belief (Flavell and Miller, 2002). Nonverbal experiments have suggested that even 15-month-old infants have an understanding of false belief (Onishi and Baillargeon, 2005), but the results are disputed (Perner and Ruffman, 2005). Even if Onishi and Baillargeon are right that 15-month-olds have some kind of understanding of false belief, otherwise verbal 3-year-olds are unable to use this understanding when making statements about another person's beliefs. If there is an understanding of false belief, it must be quite different from the adult theory of mind.

The standard false belief task involves first order beliefs. Perner and Wimmer (1985) conducted a study of second order false belief comprehension. Children heard a story about an ice cream van. To correctly answer the questions, they needed to represent a second order belief of the following form: "John (wrongly) thinks that Mary thinks that..." Perner and Wimmer found that children from age 6 or 7 onward are able to do this. A version of this test administered to Dutch school children found that 90% of 7-year-olds succeed, but that 60% of 6-year-olds still fail (Muris et al., 1999). Hogrefe

and Wimmer (1986) investigated the relation between false belief and ignorance. They added a question to Perner and Wimmer's story of the form: "Does John know that Mary thinks that...?" About half of all 5-year-olds were able to answer this question correctly, even though most of them could not answer the subsequent false belief question. A similar lag between understanding of ignorance and of false belief was found in first order reasoning. Sullivan, Zaitchik, and Tager-Flusberg (1994) created a new second order story with less complexity, again using an ignorance question before the false belief-question. They also included more probe questions and feedback. They found that with their new story at age 5;6[1] 90% of the children could answer the false belief-question and justify their response. Even 40% of preschoolers succeeded. The experimenters conclude that processing demands are an important factor in second order tasks. They suggest that once children understand the representational nature of mental states, no further conceptual development is needed to recursively embed mental states: as long as the information-processing load is not too high, children can achieve second order reasoning (see also section 2.1.4 for more discussion on this).

Adult achievements

Adults correctly use second order reasoning in the experiments just mentioned. Higher orders of reasoning are very difficult because they place large demands on working memory. Keysar, Lin, and Barr (2003) report that even first order reasoning is not used as a part of routine operation. They conducted an experiment in which adult subjects had to follow instructions for moving objects across a grid. The instructor could not see some hidden objects that the test subject knew about. Nevertheless, 71% of the test subject interpreted the instructions at least once as referring to a hidden object, and tried to move it. 46% of the test subjects did so most of the time. This study demonstrates that ToM reasoning is not routinely used to infer the intentions of other people. The Keysar et al. experiment is an *applied* task, and subjects have to use their theory of mind *spontaneously* to correctly perform the task. The false belief task on the other hand asks very explicit questions about another person's beliefs; it is difficult to answer such a question without reflecting on the other person's beliefs. The Keysar et al. experiment is interesting because the (spontaneous) application of ToM that is required in this task may be closer to real life situations than the explicit questions in a false belief task. Other examples of imperfect performance in applied tasks can be found in section 2.3.2 on games in Theory of Mind research.

[1] I will follow the convention of specifying ages in years and months, with the year before the semi-colon and the month after it.

2.1.3 Some applications of ToM

The possession of ToM is of tremendous importance for social cognition and behaviour. Below I will list some less obvious domains for which ToM has been claimed to be important.

Learning

Tomasello, Kruger, and Ratner (1993) distinguish three types of learning which are essential to the invention and transmission of human culture. They claim that specific ToM-developments are necessary requirements for these types of learning. The ability for imitative learning, which they distinguish from mere emulation, arises at 9 months and requires that the learner can establish joint attention. 4 year olds can participate in instructed learning, for which first order ToM is a requirement. At age 6 or 7, the capacity for second order reasoning enables children to engage in collaborative learning.

Pragmatic reasoning in the domain of language

Pragmatics is the area of language that deals with the differences between literal sentence meaning and the speaker's meaning. The pragmatic meaning of an utterance depends on the context. Pragmatic inferences are not absolute; they can be overruled by new information. Crucially, the pragmatic meaning depends on the speaker's and hearer's prior knowledge and expectations. Speakers must take into account hearers' knowledge and hearers must take into account speakers' knowledge. Therefore, to reconstruct the speaker's meaning requires recursive reasoning about the speaker's knowledge and intentions. The hypothesis in this thesis is that this reasoning requires ToM. This topic will be described in more detail in section 2.2.

Playing strategic games

Both 'real-life' board games and games in the more abstract, mathematical sense often involve reasoning about one's opponent. The player must recognize that the opponent has intentions and goals different from one's own. Players who are able to accurately predict their opponent's action are in a better position to choose their own actions. The interesting thing about games is that the mental model about one's opponent may be recursive. After all, the opponent is building a mental model about *his* opponent (the player himself in a two player game) as well. Chess is a good example of a game in which good players use ToM: "My opponent wants to take my queen, but he knows I do not want to lose it unless (maybe) I can take his queen." Most board games are too complex for research purposes, but researchers can construct special games to measure the application of ToM.

2.1.4 Alternatives for ToM-explanations

When I claim that a child has a theory of mind, I claim that the child has a certain body of knowledge with characteristics of a theory. It is this conceptual knowledge that is responsible for the child's abilities in typical ToM-tasks. However, changes in the experimental design can have great influence on the results, i.e. on the age at which children can accomplish the task. Explanations for such differences include: ability to understand the questions (especially nested language constructs), ability to remember the story, and ability to process all the (conflicting) beliefs simultaneously. General cognitive skills and information-processing abilities are needed to succeed in many ToM-tasks. It remains a point of discussion between researchers to what extent milestones in ToM development are the result of conceptual knowledge (the mental model of others) or of information-processing abilities. Flavell (2002) is very skeptical of the idea of a 'magical transition' at age 4, and is a proponent of a more gradual view. He thinks 3-year-old's problems with the false belief task may be caused by a general representational inflexibility and an inability to inhibit certain behaviour; both problems that are not specific to ToM-reasoning. The 'windows' task (Russell et al., 1991), a deception task that does not require a representational theory of mind, showed that 3 year olds do not yet have the required cognitive skills (in this case inhibition of a prepotent response) that would be needed to succeed at a false belief task. However these results were not replicated by Samuels, Brooks, and Frye (1996). On the contrary, Samuels et al. found that 3 year old subjects succeeded on the windows task in all variations, while failing a first order false belief task. These results exclude at least one non-conceptual reason for failure on the false belief task, so lend credibility to the idea that a conceptual leap has taken place when a child succeeds at the task.

The same discussion (conceptual knowledge versus information-processing) can be held with regard to the leap from first to second order reasoning. Sullivan et al. (1994) are proponents of the idea that no new conceptual knowledge is involved in the transition from first to second order reasoning. Although little research has been done on higher than second order reasoning, it is generally believed that the problems adults experience with this kind of reasoning are due to information-processing limitations (especially working memory) rather than conceptual differences.

This thesis will not explore the questions described above any further. I assume that ToM reasoning involves a conceptual component, but this need not be the only factor involved. When a greater variety of ToM reasoning tasks will have been developed, correlations between those tasks and with more general cognitive tasks may reveal how important the conceptual component is. At the present moment this question can't be answered. But I will of course consider the possible confounding effect of information-processing and language ability in designing my experiments.

2.2 Reasoning about speaker's alternatives

In this chapter I discuss several phenomena in language in which the listener must reason about the speaker's alternatives to correctly interpret the sentence. These phenomena also have in common that children acquire the correct interpretation relatively late.

2.2.1 Scalar Implicatures

A much studied phenomenon in pragmatics is the phenomenon of scalar implicatures. Scalar implicatures are a special kind of conversational implicatures. Below is an example from Papafragou and Tantalou (2004):

A: Do you like California wines?
B: I like some of them.
A can now conclude that B does not like all California wines.

In this example, the term 'some' is used to communicate 'some but not all'. It is called a scalar implicature because the terms 'some' and 'all' can be placed on a scale from least informative to most informative. The semantic meaning of 'some' is 'at least one'. If it were the case that B liked all California wines, both 'some' and 'all' would have been semantically correct terms to describe the situation. The term 'all' however is more informative than 'some'. A uses the fact that the speaker did not choose this more informative term to conclude that the informative term was not appropriate for the situation, and therefore concludes 'some but not all'. I will call 'some but not all' the pragmatic meaning of 'some', while the logical meaning is 'some, possibly all'.

An explanation for these inferences can be given using Grice's Quantity Maxim, which can be summarized as: "Be as informative as is required for the current purpose of the exchange, yet do not be more informative than is necessary." The first clause of this sentence is also called the Q-principle, while the second clause is called the I-principle. If the speaker had liked all California wines, his use of the term 'some' would have been a violation of the Q-principle. The hearer needs to consider all this to arrive at the intended meaning. He must reason about the speaker's alternatives and about his communicative intentions - a kind of reasoning that resembles ToM.

Several studies have been carried out to investigate children's acquisition of the pragmatic meaning of 'some'. Noveck (2001) tested children using infelicitous sentences. An example of an infelicitous sentence is: "Some elephants have trunks." The sentence is logically correct, but infelicitous because all elephants have trunks. Noveck found that 89% of 7/8-year-olds and 85% of 10/11-year olds gave logical responses (i.e. they agreed with the sentence), but only 41% of adults did. Papafragou and Musolino (2003) found similar results with 5-year-olds, but they were able to increase the pragmatic response rate to 52.5% by training the children on a similar task and providing more context. A different

study by Papafragou and Tantalou (Papafragou and Tantalou, 2004) with children aged 4-6 found that 70-90% were able to make the pragmatic implicature. Feeney, Scrafton, Duckworth, and Handley (Feeney et al., 2004) found no difference between 8-year-olds and adults, but it should be noted that in both groups there was a high proportion of logical (instead of pragmatical) answers.

How can these results be interpreted? A problem with scalar implicatures is their defeasibility. A hearer who suspects that the speaker is uncooperative, stupid (in a typical experiment, the speaker will also utter blatantly false sentences such as "All birds have fur"), or just uninformative, will not make the implicature. This accounts for the variability of adult responses in these studies. Still, young children are more logical than adults. There are not just two stages, but three (Feeney et al., 2004):

1. logical interpretation only
2. pragmatical interpretation only
3. a logical interpretation that results from choice. The pragmatic interpretation can be suppressed if the context demands this.

The defeasibility of implicatures makes it difficult to make bold claims about a person's position on this ladder of stages. Does he not make the implicature because he is unable to, or because he knows about its defeasibility? Therefore I will now turn my attention to other linguistic constructs that are acquired late in childhood, but for which the adult interpretation is less ambiguous. I will need the framework of Optimality Theory to explain them.

2.2.2 Optimality theory

Optimality Theory (OT) is a linguistic model proposed by Alan Prince and Paul Smolensky in 1993. Its main idea is that the forms of language arise from the resolution of conflicts between violable constraints. The candidate that 'wins' is the form that incurs the least serious violations, determined by the hierarchy of constraints. Constraints are universal, but their hierarchy differs from language to language. When speaking, the 'input' of the process is the intended meaning and the candidate 'outputs' are the possible forms. When listening, the input is the form and the candidates are the possible meanings. In both processes the same constraints are used, but not all constraints always apply. Faithfulness constraints require that the output matches the input in some way. Markedness constraints impose requirements on the output, without regard for the input. For more information on optimality theory, see Blutner, De Hoop, and Hendriks (2006).

Candidates and their constraints are usually listed in a tableau to determine the winner. For the simple cases discussed in this thesis no tableaux are necessary. I will use the $>$ (greater than) and $<$ (smaller than) signs to indicate that some form-meaning pairs are more or less optimal than others.

The kind of optimization discussed so far can be characterized as one-dimensional and unidirectional. The hearer hears a certain form and selects the best meaning from a set of candidates. Blutner (Blutner, 2000) proposed a two-dimensional, bidirectional framework. In this two-dimensional framework the Q- and I-principles are formulated in the following way:

Strong optimality
A form-meaning pair (f,m) is optimal iff[2] it satisfies both the Q- and the I-principle, where:
(Q) (f,m) satisfies the Q-principle iff there is no other pair (f',m) such that (f',m) > (f,m)
(I) (f,m) satisfies the I-principle iff there is no other pair (f,m') such that (f,m') > (f,m)

At first sight it looks as if only the I-principle is applicable to the hearer's interpretation process, because only this principle compares different meanings for a given form. But in the two-dimensional framework, the hearer is no longer just comparing candidate meanings, but comparing form-meaning pairs. The hearer uses both principles to evaluate form-meaning pairs, and eliminates ('blocks') the pairs that are not optimal. Since the Q-principle is a comparison of different forms, the hearer is taking the perspective of the speaker to compare different alternatives. Of the form-meaning pairs that remain, the pair whose form corresponds to what was actually heard will determine the interpretation.

I will use the pronoun interpretation problem analysed by Hendriks and Spenader (Hendriks and Spenader 2004; Hendriks and Spenader, to appear) to illustrate the bidirectional framework. Consider the interpretation of the object in the following two sentences:

(1) The boy saw himself.

(2) The boy saw him.

There are two different forms: the reflexive used in sentence 1 and the personal pronoun used in sentence 2. There are also two interpretations possible: the object may *corefer* with the subject to the same boy, or the object may refer to some other person *disjoint* from the subject. These forms and meanings can combine to form 4 form-meaning pairs: (pronoun, disjoint); (pronoun, coreferential); (reflexive, disjoint); (reflexive, coreferential). I will abbreviate these terms as 'reflex' (reflexive), 'coref' (coreferential), and 'disj' (disjoint). Two constraints are assumed. The first constraint says that a reflexive must be bound locally and gives rise to the ranking (reflex, coref) > (reflex, disj). The second constraints prefers reflexives over pronouns and gives rise to the rankings (reflex, coref) > (pronoun, coref) and (reflex, disj) > (pronoun, disj). With these rankings, the form-meaning pairs can than be ordered as below:

[2] if and only if

(reflex, coref) ⟵ (pronoun, coref)
↑
(reflex, disj) ⟵ (pronoun, disj)

The horizontal arrows in this diagram indicate that the left pairs are better than the right pairs by the Q-principle. The vertical arrow indicates that the upper pair is better than the lower pair by the I-principle. Since all pairs but the upper left one have arrows departing from them, only the upper left pair is optimal. This optimal pair corresponds to the preferred interpretation of sentence 1. Unfortunately, strong optimality can not explain the existence or interpretation of sentence 2.

There is however an alternative mechanism called 'weak optimality', formulated in the following way:

Weak optimality
A form-meaning pair (f,m) is optimal iff it satisfies both the Q- and the I-principle, where:
(Q) (f,m) satisfies the Q-principle iff there is no other pair (f',m) such that (f',m) > (f,m) such that (f',m) satisfies I.
(I) (f,m) satisfies the I-principle iff there is no other pair (f,m') such that (f,m') > (f,m) such that (f,m') satisfies Q.

In this variant the pair (reflex, coref) is still optimal. The pair (pronoun, disj) becomes optimal as well. Although (reflex,disj) is better than (pronoun,disj) by the Q-principle, (reflex,disj) itself does not satisfy the I-principle because (reflex, coref) is better. Even if there were an arrow indicating that (pronoun, coref) is better than (pronoun, disj) by the I-principle (there is no such arrow with our current choice of constraints), this would not make (pronoun, disj) suboptimal because (pronoun, coref) does not satisfy the Q-principle. Therefore, the pair (pronoun,disj) is weakly optimal.

There are now two optimal pairs: (pronoun, disj) and (reflex, coref). These two optimal pairs correspond with the preferred interpretations of sentence 2 and 1 respectively. The beauty of this approach is that only two constraints are needed to arrive to arrive at this ordering. Other explanations of this phenomenon typically involve a third constraint or principle. Furthermore, this problem is interesting because children up to age 6 interpret personal pronouns as coreferring with the subject about half the time, despite evidence from production data that they have competence in the relevant constraints. (Bloom et al., 1994; cited in Hendriks and Spenader, 2004). If the two constraints are applied in a *unidirectional* framework, children's correct production and incorrect comprehension can be explained. The explanation for children's aberrant interpretation of reflexives is not that they have different constraints than adults (the production data contradicts this), but that they do not apply bidirectional optimization.

2.2.3 The markedness principle

In the pronoun interpretation problem, it may not be obvious that a pragmatic inference is drawn, because both forms are very prevalent and both interpretations readily available. Yet the bidirectional mechanism is sufficient to explain the phenomenon without having to 'fix' the preferred pairs in the grammar or in the vocabulary. The pragmatic aspect is more discernible in pairs where one of the forms or meanings is *unmarked* and the other *marked*. An unmarked form or meaning is a property that is widespread across languages, occurs often, is neutral (meaning), does not have overt morphological marking (form), and is acquired earlier. A marked form or meaning is the exact opposite: it doesn't occur very often, it is somehow 'special'. The markedness principle states that unmarked forms receive unmarked interpretations, while marked forms receive marked interpretations. Bidirectional optimality can explain how these form-meaning pairs arise simply by assuming two constraints such that *unmarked form* > *marked form* and *unmarked meaning* > *marked meaning*. These constraints result in the following ordering:

(unmarked form, unmarked meaning)	⟵	(marked form, unmarked meaning)
↑		↑
(unmarked form, marked meaning)	⟵	(marked form, marked meaning)

Again we can conclude from the constraints that the pair (marked form, marked meaning) is strongly optimal. The pair (unmarked form, unmarked meaning) is weakly optimal: although (unmarked form, marked meaning) is better by the Q-principle, it does not itself satisfy the I-principle. And although (marked form, unmarked meaning) is better by the I-principle, it does not itself satisfy the Q-principle.

What happens if one does not use bidirectional optimization? In that case a speaker will simply compare the two possible forms and always choose the unmarked form (because unmarked form > marked form). Similarly, a listener will compare the two possible interpretations and choose the unmarked interpretation. Thus we would expect young children to always choose an unmarked interpretation even if a marked form was heard, and always produce unmarked forms even if a marked meaning was intended. This is different from the pronoun interpretation problem, in which the constraints are such that adultlike production for both intended meanings can be achieved without bidirectionality.

The markedness principle can be applied to well-defined grammatical constructions, but it can also be creatively applied to create a potentially infinite number of special, marked sentences. Van Rooij (Van Rooij, 2004) gives the following sentence: "Miss X produced a series of sounds that corresponds closely with the score of 'Home Sweet Home' ". The long form "produced a series of sounds that corresponds closely with the score of" is a marked alternative to the word "sang". The longer, suboptimal form is used to indicate a special meaning: there was something wrong with Miss X's singing. Unlike in the pronoun interpretation problem, the marked meaning is not precisely specified: we don't know exactly how Miss X's singing was different from ordinary singing, but we know the

interpretation must somehow be marked and special. Steerneman et al. (2003) regard the understanding of complex humour and irony as an advanced social-cognitive skill, succeeding the acquisition of second order reasoning. However, I am not aware of a study that investigates only those ironic sentences that can be explained by the markedness principle.

2.2.4 Indefinite objects and subjects in Dutch

I will apply the markedness principle to the interpretation of indefinite subjects and objects in Dutch, as described by De Hoop and Krämer (2006). The example sentences in this section are copied from their article.

Consider the following two sentences with an indefinite object noun phrase:

(3) Je mag twee keer een potje omdraaien.
You may two time a pot turn-around.
"You may turn a pot twice."

(4) Je mag een potje twee keer omdraaien.
You may a pot two time turn-around.
"You may turn a pot twice."

The adult interpretation of sentence 3 allows for two different pots to be turned, which is called a *non-referential* reading because the indefinite object "een potje" does not refer to a specific pot in the world. Sentence 4 on the other hand receives a *referential* reading: the command is only executed correctly if the same pot is turned twice. Most children below age 7 will interpret *both* sentences non-referentially.

Young children's preference for a non-referential reading of the indefinite is specific to objects. Subjects *do* receive a referential reading. Take for example this sentence:

(5) Een meisje gleed twee keer uit.
A girl slipped two time out.
"A girl slipped twice."

Children as well as adults interpret sentence 5 as "A certain girl slipped twice" and not as the non-referential "Twice a girl slipped." Unlike the English translation, sentence 5 is not ambiguous in Dutch. The non-referential meaning can be expressed as follows:

(6) Er gleed twee keer een meisje uit.
There slipped two times a girl out.
"Twice a girl slipped."

But while adults assign a non-referential reading to the subject "een meisje", children do not. Until age 10 the majority of children prefer a non-adultlike referential interpretation. In conclusion, children treat all objects as non-referential and all subjects as referential, without regard for the word order of the sentence. Their preference matches the general cross-linguistic pattern that subjects are *usually* referential and definite while objects are *usually* non-referential and indefinite. Although Dutch has specific, unambiguous forms to express referential and non-referential meanings, children are unable to depart from the general pattern.

Both the adult interpretation and children's failure can be explained using the markedness principle.

The position of the object in sentence 3, to the right of the adverbial phrase, is the most common position. It can therefore be regarded as an *unmarked* form. The object position in sentence 4 is called *scrambled*. It is a special, less common and therefore *marked* form. For objects, the non-referential meaning is the *unmarked* meaning, in accordance with cross-linguistical pattern that objects usually non-referential. The referential meaning is therefore the *marked* meaning. Marked forms receive marked interpretations while unmarked forms receive unmarked interpretations. Thus, sentence 3 has an unmarked, non-referential meaning while sentence 4 has a marked, referential meaning. We saw that young children assign a non-referential (unmarked) interpretation to both sentences. This can be explained by assuming that they do not use bidirectional optimization. We would also predict that these children will not produce marked forms to express a marked meaning.

For the subject sentences, sentence 5, with the subject in the standard, sentence-initial position, is the unmarked form. Sentence 6, using the existential construction with the extra morpheme 'er' (there), is the marked form. For subjects it is the referential interpretation that is most common and therefore unmarked. Again, the adult interpretation can be explained from the principle that unmarked forms have unmarked meanings and marked forms have marked meanings, while children choose the unmarked form or meaning because they do not apply bidirectional optimization.

2.2.5 Optimality theory and ToM

In section 2.2.1 an intuitive account was given of why scalar implicatures require ToM-reasoning. The same applies to those phenomena that I analysed using optimality theory. I will now attempt a more intuitive account of the bidirectional mechanism to illustrate how ToM-reasoning is used. Consider a person Peter who hears sentence 6 uttered by Sally:

(6) Er gleed twee keer een meisje uit.
"Twice a girl slipped."

This sentence is in the marked (existential) form. Peter considers that Sally could have meant the unmarked, referential meaning. But he also knows that, if Sally had intended this, she would have used sentence 5 (by the Q-principle). And he knows that Sally knows that, if she used sentence 5, Peter would assign an unmarked interpretation to it. Therefore Sally must have used the marked form on purpose: she wants Peter to have a non-referential interpretation to it. Peter's reasoning involves two steps: first about Sally's alternatives, and then about his own interpretation of these alternatives. There is a lot of similarity with second order ToM: Peter has a model of Sally's alternatives (possible intentions) but also of Sally's model of his own interpretations (possible beliefs) of those alternatives.

Earlier I claimed that bidirectional optimization is a sufficient explanation for the (strongly and weakly optimal) form-meaning pairs that we observe and that it is not *necessary* to store these form-meaning pairs. But this does not mean that no form-meaning pairs are ever stored. It seems likely to me that common form-meaning pairs, such as the interpretation of personal pronouns and reflexives, are stored, since deciding that a pronoun does not refer to the subject of the sentence does not require nearly as much effort as answering a second order false belief question. But the pattern in which children develop the correct interpretation leads me to believe that bidirectional optimization does play a role in acquisition. Van Rooij (2004) offers a plausible mechanism of how form-meaning pairs, or even the markedness principle itself, may have evolved. This account shows that suboptimal form-meaning pairs (marked pairs with marked meanings) can evolve even if no individual language users apply bidirectional optimization. It is not clear if this means that Van Rooij thinks individual language users do not use bidirectional optimization at all.

2.3 Game theory

2.3.1 Concepts of game theory

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 (or at least independently from the other player). 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 (one for each 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 start 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 several, sequential moves. Although a sequential game may be specified in normal form by dedicating a row or column to each possible combination of moves for that player, the normal form does not specify how a player makes a single move: it just assumes that each player picks a fully specified strategy (contingent upon the other player's moves) in advance of the game. 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 Nash equilibrium is not necessarily *Pareto optimal*. A solution (or matrix cell) is Pareto optimal if there is no other solution (anywhere in the matrix) that would be preferred, or not opposed, by any other player.

A *zero-sum* game is a game in which the sum of the payoffs in each cell is zero. This means that one player's gain is another player's loss. Optimal strategies for zero-sum games can be computed by the minimax algorithm (for an implementation see Russell and Norvig, 1995). Nash equilibria in zero-sum games are always Pareto optimal, since it is never possible to increase the payoff of one player without decreasing another player's payoff. *Non zero-sum* games present more difficulty in analysing. In the famous prisoner's dilemma game, which is non-zero, the Nash equilibrium of mutual defection is not Pareto optimal.

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.

2.3.2 Games in Theory of Mind research

Games can be designed so that they require ToM for optimal performance. The use of games for ToM-research has a number of advantages. First, games are different from the false belief task in that they do not depend on language skills very much. Games are interesting because they are *applied* tasks. Using ToM gives the subject some advantage in the game, but the subject is not explicitly asked to use ToM. We saw from the Keysar et al. (2003) article 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 row-column game. Although the article does not explicitly discuss ToM or order of reasoning, my analysis of it will. 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. Perner also identified a number of individualistic strategies that did not take the opponent into account. These strategies were: choose the row that contains the cell with the highest payoff (maximax); choose the row with the highest average payoff (maxaverage); and: avoid the row that contains the lowest payoff (maximin). Perner made sure that for two of the three matrix types, none of these individualistic strategies could result in selecting the optimal row, so that a correct prediction was really necessary. Therefore, this research measures first order ToM-reasoning.

All children played both as column and as row, and half of the children were asked to predict 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) then 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 action and predictions are crossed there are four possible outcomes. The most common outcome (40%) was that children focused on the cell that contained their highest payoff: they chose the row that contained this cell and they predicted that the opponent would choose the column that contained this cell. It seems that these young children were unable to take their opponent's point of view, even though it would have helped them. Older children were able to make correct predictions: when playing as row about 74% of all prediction 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. I think this 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. They 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 subjects 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 chapter 4.

2.3.3 Game theory and optimality theory

Dekker and Van Rooij (2000) show how bidirectional optimality can be analysed as an application of game theory. The four possible form-meaning pairs can be arranged in a 2 by 2 matrix in normal form. It is then easy to show that there can be two Nash equilibria. The Nash equilibrium corresponding to the strongly optimal pair is also Pareto optimal; the Nash equilibrium corresponding to the weakly optimal pair is not. Although Dekker and Van Rooij show an interesting correspondence between optimal pairs and equilibria in game theory, using their game representation to show how listeners interpret a sentence is problematic (Van Rooij, 2004), since this is a sequential process that can not be described very well in normal form.

Chapter 3

The research question refined

I started my work on this research with the following research question:

> How does children's development of the ability to reason about other people's knowledge and intentions correlate with the development of the ability to reason about speaker's alternatives during language comprehension?

This question is not precise enough to guide experimental research. It couldn't possibly have been, because it was conceived before the literature study was completed. In this chapter I will first summarize the answers to this question that can be found in the literature. After this, new and more precise questions will be formulated to guide the experimental part of this project.

3.1 How does reasoning about other people's knowledge and intentions develop?

The standard tool to investigate ToM-reasoning is the false belief task. Children succeed at a first order false belief task at 4 years of age. Second order false belief develops later: at 6/7 years. Applying ToM-reasoning to more practical tasks such as games is more difficult than ToM-reasoning in a false belief task. Even adults do not achieve perfection. No experiments with applied ToM-tasks have been conducted on both adults and children, so it is unknown whether applied ToM-reasoning continues to develop after age 6/7. Different applied tasks have not been correlated with each other or with a standard false belief task, so it is unknown if these tests provide a good measure of ToM-reasoning. An advantage of game-like applied tasks is that more differences between individuals can be measured than with a false belief task.

3.2 How does reasoning about speaker's alternatives develop?

In the preceding chapter the following linguistic phenomena were studied:

- *Scalar implicatures.* The word 'some' often carries the implicature 'not all'. Since 'all' is not really a variant form of 'some', this phenomenon can not easily be described with the OT framework, but principles similar to bidirectional optimization apply. Age of acquisition: 4-7 year.
- *Interpretation of pronouns.* Reflexive objects corefer with the sentence subject, personal pronouns receive a disjoint interpretation. This requires bidirectional optimization with weak optimality (strong optimality is not sufficient). Age of acquisition: 6 and up.
- *'Ironic' sentences* like the sentence about Mary's singing. Steerneman et al. (2003) regard the understanding of complex humour and irony as an advanced social-cognitive skill, succeeding the acquisition of second order reasoning.
- *Indefinite objects.* The default interpretation is non-referential. The marked form receives a referential reading, acquired around age 6.
- *Indefinite subjects.* The default interpretation is referential. The existential form receives a non-referential interpretation, which in still difficult to many children up to and including age 9.

A plausible explanation was given for why the correct comprehension of these sentences involves second order theory of mind. The late age at which the correct interpretations are acquired is consistent with this idea. However, there is no *within subjects* experimental work that investigates whether there is a link between these language phenomena and theory of mind reasoning. Moreover, the different constructions are not acquired at the same time. Although it is possible that second order theory of mind must precede acquisition of all these linguistic constructions, an additional factor (such as linguistic experience and the commonality of the frequency of these constructions in the language) is needed to account for the fact that some constructions are acquired earlier than others.

3.3 How do these developments correlate?

There is no experimental research about a possible correlation between ToM-reasoning and reasoning about speaker's alternatives. For most of the language experiments I studied the adult interpretation is acquired at an age of 6 years or later. This seems to match the age of acquisition of second order reasoning. However, such age comparisons between

different experiments constitute only very weak evidence. A within subjects experiment is desired.

3.4 The refined research question

The original research question is about a correlation between two abilities: theory of mind reasoning and reasoning about speaker's alternatives. If reasoning about speaker's alternatives involves theory of mind, it is an applied ToM-task. Therefore it would be best to try and correlate it with another applied ToM-task. I have selected Hedden and Zhang's (2002) strategic game for this.

Given how little research there has been on applied ToM-tasks, especially second order tasks, I think it is best that a standard second order false belief task is also conducted to investigate the relation between these two different measurements of ToM. Since no experiment has compared children and adults on a second order applied ToM-task before, this research is in itself valuable even without linking it to language development. So in the refined research question below I am no longer relating *two* things, but *three*: false belief, applied ToM reasoning, and language comprehension.

> How does second order reasoning develop and how is it applied to strategic games and reasoning about speaker's alternatives?
>
> 1. Is performance on a false belief task related to performance on a strategic game?
> 2. Is performance on a sentence comprehension task in which correct comprehension requires reasoning about the speaker's alternatives related to performance on other ToM-tasks (the false belief task or the strategic game)?
> 3. What are the differences between adults and children for all three tasks?

I am aware that 'related' is an imprecise term, but I have used it because the term correlation is too restrictive. Within subjects correlations between the tasks are of course most desirable, but what if performance on a certain task is (nearly) perfect in one or both age groups? In that case it is not possible (for lack of variation) to find correlations with other tasks in that age group, but it may still be possible to conclude that mastery of one task precedes mastery of another task.

Chapter 4

Design

This project investigates how ToM-reasoning is applied to strategic games and to the comprehension of sentences that can be explained with bidirectional OT. My experiment will use a within subjects design, in which all subjects participate in three tests:

- a standard second order false belief task
- a sentence comprehension test on indefinite subjects
- a strategic game, based on the game by Hedden and Zhang (2002)

There will be two groups of subjects: children from 'groep 5' (age 8-10 years) and adults. According to Vrieling (2006) and De Hoop and Krämer (to appear) some children in this age group still have the child-like interpretation of existential sentences, while others already have an adultlike interpretation. Thus, this age group will provide sufficient variation between subjects in this task to prove or disprove correlations with other tasks.

I expect that performance on the second order false belief task will be near perfect, since the majority of American children are able to succeed at this task at the age of 6 (Perner and Wimmer, 1985). Succeeding at such an explicit second order task is a necessary condition for succeeding at other tasks requiring second order ToM, but it may precede them. After all, even adults do not always apply their second order reasoning skills when needed. The strategic matrix game is therefore included to measure the *application* of second order reasoning. The three tests were always administered in the same order and in one session, which is the order in which I describe them.

The next sections will first describe the subjects and then the design of each of the three tests.

4.1 Subjects

There were two groups of subjects: adults and children from 'groep 5' (age 8-10 years). Each subject participated in three tests: the strategic game, the sentence comprehension test, and the false belief test (in that order). In the language test, items were balanced so that half of the subject received first an existential sentence (like sentence 6 in section 2.2.4) and then a canonical sentence (like sentence 5 in section 2.2.4) , the other half first a canonical sentence and then an existential sentence. For the child group, the order of the stories in the false belief test was also balanced. The tests were administered in one session that took about 30 minutes. The strategic game was played on a laptop computer with a separate mouse.

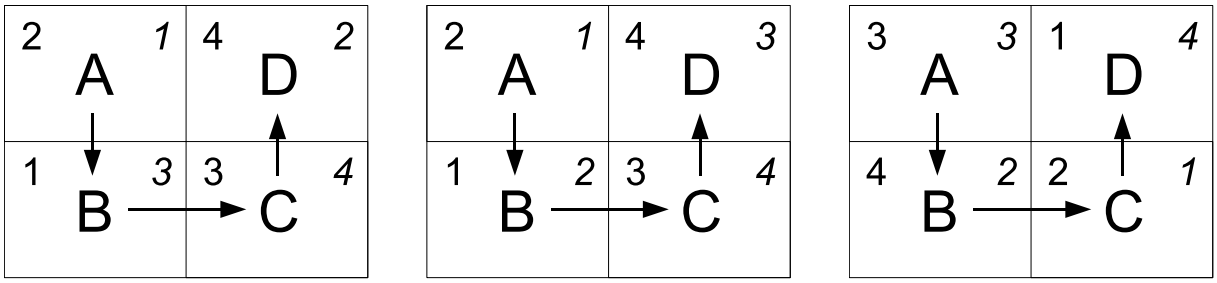
The adult subjects were psychology students participating for course credit. There were initially 31 subjects, but the first four subjects were excluded because of the subsequent change in the reward structure and instruction of the strategic game, which will be described in section 4.2.2. Of the remaining 27 subjects, 10 were males and 17 were females The youngest was 18 and the oldest was 26 (median age 20 years). Two subjects who were not native Dutch speakers were excluded from the language test. Two other subjects reported being bilingual in Dutch and Frisian. All other subjects were monolingual Dutch speakers.

The child subjects were 40 children from 'groep 5' from the St. Jorisschool in Heumen and the Christelijke Basisschool de Bron in Marum, 21 girls and 19 boys. The youngest was 8;4 and the oldest was 10;3; the mean and median age were both 9;2. All children were native Dutch speakers.

4.2 Strategic game

In the previous chapter I mentioned Hedden and Zhang's (2002) matrix game as an example of a strategic game to measure the application of second order ToM. It is the only applied task I could find that is designed to distinguish first and second order ToM. The game allows for a lot of repetition, so that differences in performance between individuals can be measured accurately.

Before explaining my own design decisions I will first give a detailed summary and analysis of Hedden and Zhang's experiment.



(a) (b) (c)

Figure 4.1: Three matrix games, with payoffs, presented as a matrix. The player's payoffs are at the left of each cell and undecorated, the opponent's payoffs are at the right of each cell and italic.

4.2.1 Summary of Hedden and Zhang's strategic game

Game design

Hedden and Zhang investigate the mental models that adults apply in strategic matrix games. The game they use is a sequential game of perfect information. The game has 4 cells, labeled A-B-C-D, and each cell contains unique payoffs for each player. Two players alternate in choosing either 'stay' or 'switch'. A 'stay' decision terminates the game immediately and the players receive their payoffs in the current cell. A 'switch' decision changes the current cell to the next. Figure 4.1 shows three games. Although Hedden and Zhang present their cells in a 2 by 2 matrix, the game can be better thought of as a ladder or a tree, since from each cell one can only switch to the next. A tree presentation of Hedden and Zhang's game is given by Colman (2003) and also in Figure 4.2. The game terminates if the 4th cell is reached, so there are three decision points (in cells A, B, and C in the matrix presentation). The game is non-cooperative, i.e. each player tries to maximize his own payoff. The game doesn't allow communication between the two players.

Subjects' order of reasoning

In Hedden and Zhang's experiment, test subjects play the game against a computer opponent, but are made to believe they are playing against another human. The test subject is always the first to make a decision and has two decision points, while the computer opponent is second and has only one decision point. This gives the subject more power than the computer opponent. Before making his own decision, the subject is asked to predict what his opponent would decide if the second cell were reached. Thus a subject's ability to make correct predictions and his ability to make rational decisions can be assessed separately.

Hedden and Zhang distinguish three strategies:

- *Zeroth order*
  The player only takes into account his own desires and the state of the world. This means that the player compares his payoffs in cell A and B. If his payoff in cell A is larger than in cell B, he will stay, otherwise he will move. A zeroth-order player does not look at his opponent's payoffs.

- *First order*
  A first order player takes into account his opponent's desires and assumes that his opponent will act as a zeroth-order player. A first order player will therefore compare his opponent's payoffs in cell B and C to decide whether his opponent will stay in cell B or switch to cell C. The player will then compare his own payoff in cell A with his payoff in either cell B or cell C, depending on what he predicts his opponent to do.

- *Second order*
  A second order player takes into account his opponent's desires, and is even able to take into account his opponent's beliefs about his own desires: he perceives his opponent as a first order player. The first step in this strategy is therefore to decide what his opponent would expect him to do in cell C - this is a simple comparison of his own payoffs in cell C and D. Based on this, he must compare his opponent's payoff in cell B with his opponent's payoff in either cell C or cell D to decide whether his opponent will stay in cell B or switch. Finally he must compare his own payoff in cell A with the cell that he has just predicted would be reached if he were to switch.

For the example game in Figure 4.1a, a zeroth order player would choose to stay in cell A, because his payoff in cell A is larger than in cell B. A first order player would predict that, given the choice, his opponent switches from cell B to cell C, because the opponent's payoff in cell C is larger than in cell B. Since his own payoff in cell C is larger than in cell A, the player would move. A second order player would predict that his opponent does *not* switch from cell B to cell C, since the opponent should know that he will then end up in cell D which has a lower payoff than cell B. Thus, his opponent would stay in cell B, and the second order player decides to stay in cell A. Because the game has three decision points, higher orders than second order reasoning are not useful and would lead to the same result as second order reasoning. In the example game the zeroth order player and the second order player will choose the same action: stay. However in the test procedure players are forced to predict their opponent's action and it is these predictions which are used to determine a player's order of reasoning. By definition a prediction can not be zeroth order, but a zeroth order player might make random predictions. Such players must be filtered out.

Item design

Hedden and Zhang use a training block consisting of 'trivial games'. Figure 4.1b shows a trivial game. In these games, the opponent's payoff in cell B is either larger than the payoffs in both cell C and D, or smaller than both these payoffs. Therefore, a first order and a second order player should make the same predictions. The training block allows the player to learn the game without learning much about his opponent's strategy, and it allows the experimenter to find and exclude zeroth order or guessing players.

The training is followed by two test blocks. In the first test block, all games start with a payoff of 3 for the player. In the second test block the starting payoff is 2. Each test block is divided into four sets, each consisting of four diagnostic (balanced) items and one control item. Hedden and Zhang are interested in how their test subjects infer the opponent's strategy, which can be either zeroth order (myopic) or first order (predictive). The opponent may also switch strategy between the two test blocks. Only if a subject is pitted against a first order player will a second order strategy be appropriate. Since I want to measure second order reasoning ability in all my subjects, they will always play against a first order opponent in my experiments. I will focus on those results from Hedden and Zhang that are relevant to my research. I will sometimes refer to the second order predictions and actions as 'correct' and to other actions as 'incorrect'.

The game in Figure 4.1a is an example of a diagnostic game because it allows one to distinguish first and second order strategies. The control items are 'neutral': a player using Hedden and Zhang's proposed first order strategy should make the same predictions as a second order player. Figure 4.1c shows a control item in which both a first order and a second order player would predict their opponent to stay. The control items are different from the training items: in the training items the value of the opponent's payoff in cell B lies between the values in cell C and D, but this is not the case for the control items.

Hedden and Zhang report that there is no difference in performance on control items across conditions (opponent strategy and item switch). Unfortunately they do not report what this performance is. Since a first and second order strategy should yield the same prediction on the control items, we would expect that nearly 100% of predictions are consistent with second order reasoning. Any performance below 100% would indicate that a subject is using neither first order nor second order reasoning but an alternative strategy or perhaps guesswork. In Hedden and Zhang's analysis of their test items, subjects are assigned a score representing the proportion of correct second order predictions. Any prediction that is not second order is assumed to be first order. A score of 100% indicates that a subject is using second order reasoning all the time. A score of 0% indicates that a subject is using first order reasoning all the time. A score in between indicates that a subject is using first order reasoning some of the time and second order reasoning some of the time. The important assumption behind this interpretation is that a person is always using either first order or second order reasoning. Although Hedden and Zhang did ensure

(through their training) that subjects were capable of using at least first order reasoning, they do not allow me readers to test the assumption in a more direct way, by looking at the actual performance on the control items. Such a test could also indicate whether a problem with Hedden and Zhang's first order strategy, described in section 4.2.3, is serious.

Results

70 undergraduate students participated in Hedden and Zhang's experiment. 18 were excluded based on the training block, either because they made too many prediction errors or because they made too many rationality errors; these errors in the training session indicate zeroth order behaviour or guesswork. The subjects were divided into four groups based on their opponent's strategy for each test block. Prediction scores per game set were calculated in the interval [0,1], with a score of 1 corresponding to 100% second order predictions, and a score of 0 corresponding to 100% first order predictions (remember that every option the player has can be interpreted as either first order or second order). All groups start out with a prediction score of about 0.2 for the first set. The prediction scores for the myopic group remain at this value, while the prediction scores for the group with a first order opponent rise with each set. It is about 0.5 for the fourth set. For those subjects whose second test block is against a first order opponent as well, it continues to rise in the second test block and stabilizes at around 0.7.

A second experiment was conducted with another 70 subjects. In this experiment the human confederate was said to belong to either an 'intelligent' stereotype or an 'unintelligent' stereotype, while a third group of subjects were told they were playing against a computer. No difference in performance was found between these three groups. Because the opponent never switched strategies in this experiment, half of all subjects played against a first order player throughout the experiment. Again their prediction scores rose from an initial low value, this time reaching a maximum of 0.6.

From these experiments it can be concluded that people start the game with an initial zeroth order model of their opponent. Only when this model does not work do they change to a more complex, first order model. However, the average scores at the end of the experiment are far from perfect. Subjects did not always choose the action that followed rationally from their prediction: the proportion of rationality errors was about 0.25 for games starting with a payoff of 3, and about 0.1 for games starting with a payoff of 2.

4.2.2 Adaptations to Hedden and Zhang's design

In this section I describe how I adapted Hedden and Zhang's design for my own experiment. An important consideration in my design is that children should be able to play

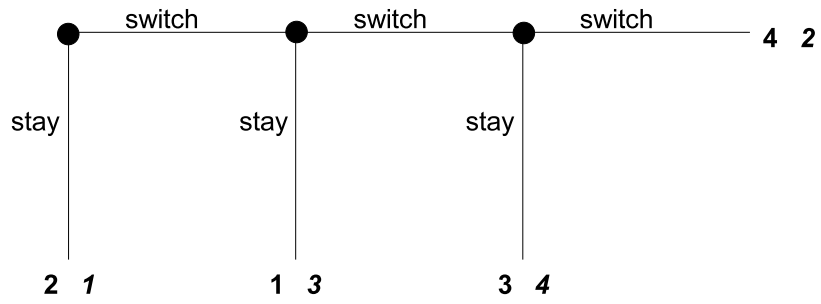


Figure 4.2: The game from Figure 4.1a presented as a tree. The player's payoffs are at the left of each cell and roman, the opponent's payoffs are at the right of each cell and italic.

the game, since both adults and children will participate in the experiment.

Game presentation

The game should be as realistic and concrete as possible, so that both adults and children will be able to play it.

I believe a board game would be more realistic than a computer game. I have piloted a board game design, but it was so slow and cumbersome that not enough items could be administered. Because of these practical problems it was decided to use a computer design for the experiment. Hedden and Zhang's second experiment showed that subjects who know they are playing against a computer do not perform better or worse than subjects who are made to believe they are playing against a human (through a computer interface). Furthermore, in a school situation where all test subjects know each other, a dyad deception would be very difficult to organize. Therefore no deception was attempted and subjects were told they would play against the computer.

Hedden and Zhang presented their game as a 2-by-2 matrix. This seems unnecessarily complex to me: although almost all cells are adjacent, only certain moves between cells are allowed. The design encourages a confusing identification between payoffs and decision points. To use the game with children, the mechanics of the game should be as clear as possible. Therefore I have chosen a 'tree' presentation (Figure 4.2). The decision points are the nodes, but the payoffs are not presented *at* the nodes, they are presented at the end of the branches.

I have chosen to use the metaphor of car driving in presenting the game. The human player and the computer opponent together control a car. The car represents the current position in the game. The decision points are represented by road junctions. The possible end points of the game are represented by dead ends. Every dead end contains a reward for each player. The rewards are colour coded: blue rewards go to the human player, yellow rewards go to the computer opponent. Rewards are displayed as a number of objects (marbles). The human player's score is presented both as a tube filled with marbles and

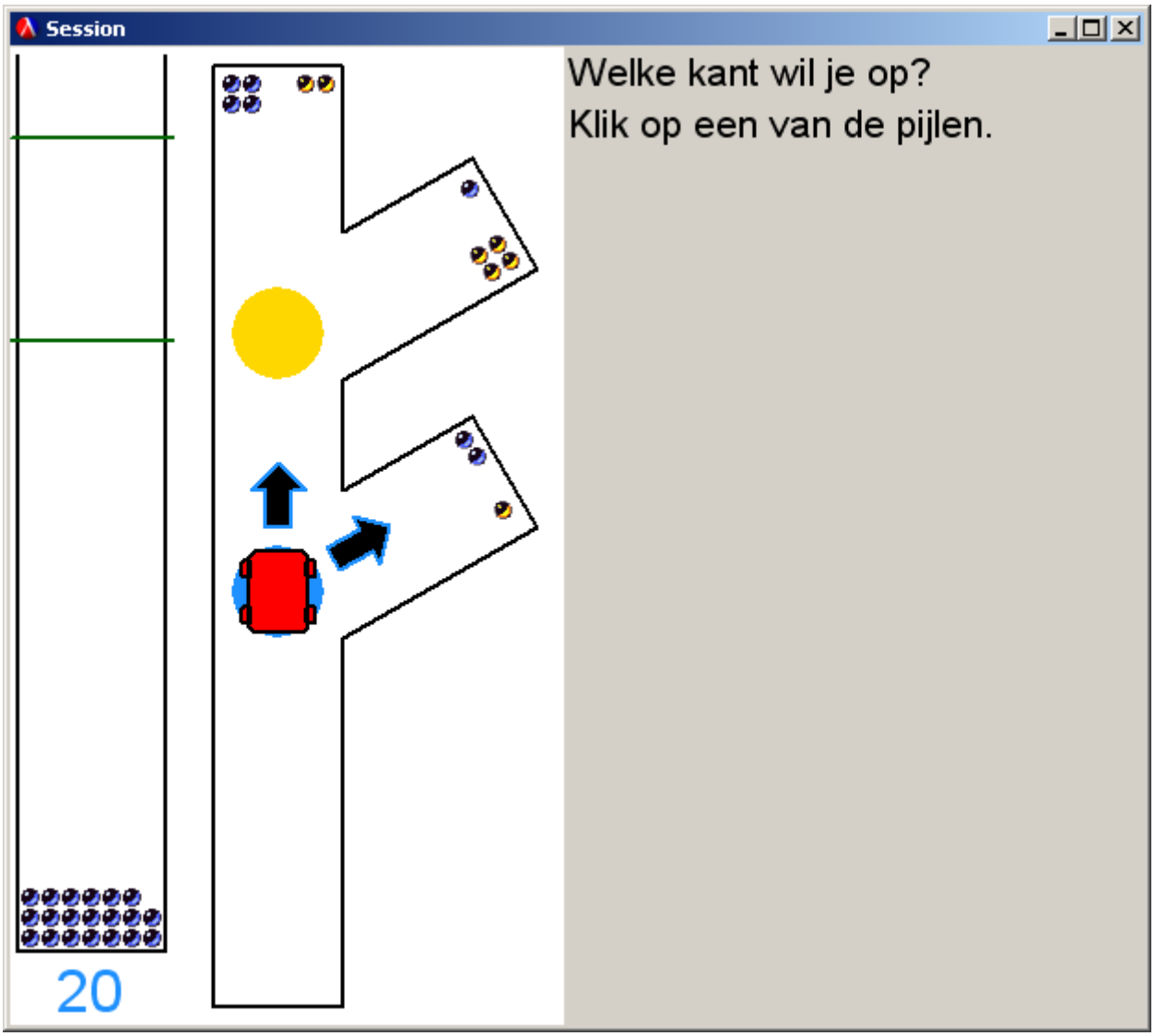


Figure 4.3: A screenshot of the computer program developed for the strategic game experiment. The tree from Figure 4.2 has been rotated to vertical. The payoffs are colour-coded and the terms 'stay' and 'switch' have been replaced with 'go straight' and 'go right'. This screenshot shows the training phase. The human player is about to choose his own action. The text at the right hand side translates as: "In which direction do you want to go? Click on one of the arrows." The tube on the left represents the player's score.

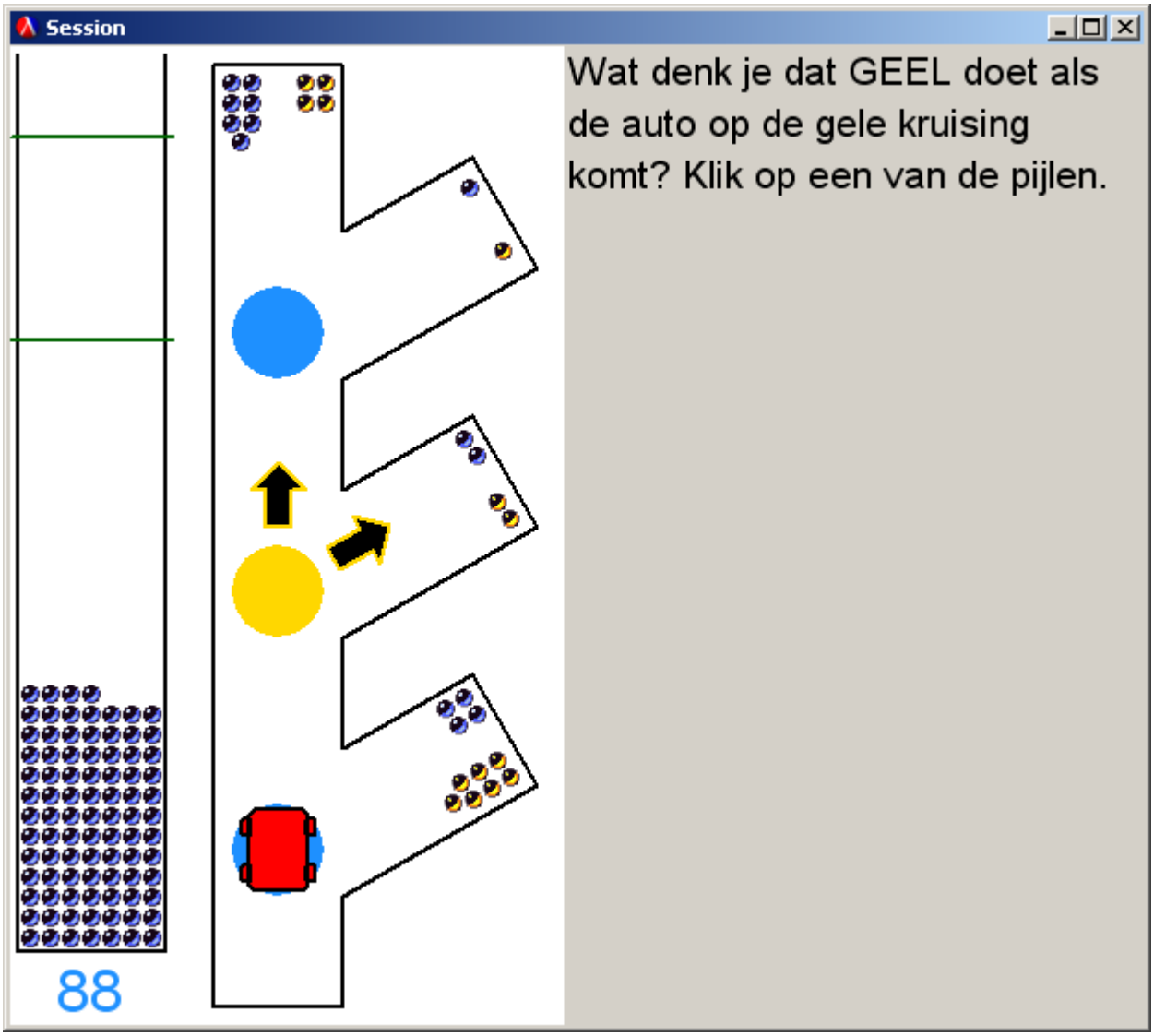


Figure 4.4: Another screenshot of the computer program developed for the strategic game experiment. This screenshot shows the testing phase. The human player is about to make a prediction. The text at the right hand side translates as: "What do you think YELLOW will do if the car reaches the yellow junction? Click on one of the arrows."

as a numeral. Each junction is marked with a colour to show which player may decide at this junction. Figure 4.3 shows a screenshot of the game in the training phase, while Figure 4.4 shows a screenshot of the game in the testing phase.

At the beginning of each game the car moves to the first junction. The player is asked to predict the opponent's action by clicking on one of the yellow arrows placed around the second (yellow) junction. After this the player chooses his own action by clicking on one of the blue arrows placed around the first (blue) junction. Depending on the action chosen, the car moves to the next junction or moves towards the 'first' reward. If the car moves to the next junction, a text message indicates which action the opponent chooses, but the action is not executed until the player acknowledges this message. If the car moves towards a reward, the reward for both players will disappear from the road and the player's reward will be added to his score. A text message will indicate how many marbles each player received. The message must be acknowledged before the next game is started. All car movements are accompanied by a sound, as is the collection of a reward. The game is played on a laptop with a 14¨ screen and a separate, optical mouse. The computer program was coded in Scheme (Abelson et al., 1998). It uses the PLT Scheme implementation, including the PLT MrEd Graphical Toolbox. The instruction for the game can be found in Appendix A.2.

Payoffs

If subjects are completely certain of their predictions and completely rational, only the relative order of the payoffs matters. To make the payoffs easier to distinguish visually and to eliminate the need for counting I will use payoffs of 1, 2, 4, and 7, rather than Hedden and Zhang's payoffs of 1, 2, 3, and 4. If subjects are not completely certain at what cell the game will end, they may weigh several possible outcomes and compare this 'aggregate' uncertain payoff to the certain payoff that will be received when the game is ended immediately. In this case the actual payoffs *do* matter. I think my alternative payoff structure is preferable, because the larger differences between the higher payoffs motivate a subject to try and reach the best cell. However one could object that this payoff structure encourages too much risk-taking behaviour.

I have been somewhat concerned that it will be difficult for subjects to disregard the opponent's score completely. Although the opponent's payoffs are useful for predicting the opponent's actions, players are not supposed to *care* what payoff their opponent receives. This may or may not be accurate when describing economic phenomena, but unfortunately, many games in people's lives (such as board games and computer games) are competitive and zero-sum, while the non-zero sum game in this experiment expects players to be only concerned with their own rewards (i.e. purely egoistic behaviour). In competitive games, players don't just maximize their own score, but simultaneously try to minimize their opponent's score. If subjects apply competitive goals to this experiment, they will not make optimal decisions. Non-representative evidence from my first four

adult subjects indicates that competitive goals may have been used, which is why the measures described in the next paragraph were adopted for the remaining subjects and why these first four subjects were excluded.

The concern that people might apply egoistic goals to Hedden and Zhang's game has already been voiced by Colman (2003). He claims that if payoffs are not tangible (preferably monetary), extraneous arguments such as the competitive goal described above can influence a player's choice. The appropriate egoistic behaviour can therefore be encouraged by making the payoffs more tangible. In my design the subject's score is prominently displayed, while the opponent's score is hidden. Also, two 'target scores' are displayed. If the subject reaches the first score, he earns a real reward. If he reaches the second score, he earns a double reward. For children the reward is a sticker, while adults receive sweets. Subjects are told in advance that they can earn rewards, and they are explicitly told that the opponent's score does not influence their rewards. If for some subjects all these precautions do not achieve the desired effect of eliminating competitive goals, they should be excluded. Competitive goals may lead to both prediction and rationality errors during the training phase. Like Hedden and Zhang these errors will be a criterion for exclusion. As an extra safeguard I will also record and analyse decisions made at the last intersection during the test phase (whenever this last intersection is reached). Since the reasoning required in this intersection is trivial and theory of mind is not involved, non-optimal decisions at this intersection justify a strong suspicion that the subject is using incorrect goals.

Item design

I will use Hedden and Zhang's diagnostic items (test items) and control items for the experiment. However, I will mix games starting with a payoff of 2 and games with a payoff of 3 (which becomes 4 in my payoff structure). This ensures a continuous (albeit possibly slower) learning progress without a special event (the change of game type) in the middle of the experiment. The set size is doubled from 5 to 10 items, so that balance within each set is maintained. Each set will consist of the items from the two corresponding sets (from both blocks) in Hedden and Zhang, in random order. All items can be found in Appendix A.1.

I think Hedden and Zhang's training has some problems, as I will now argue. It is unnecessarily complex and there may be inappropriate transfer from training to test. The goal of the training is to familiarize the player with the rules of the game, to test whether the player is capable of at least first order reasoning, and to test whether the player is acting rationally. None of this requires items with three decision points and four possible end points, as were used by Hedden and Zhang. The special payoffs required to make these complex items sufficiently simple for training may even result in inappropriate transfer. Hedden and Zhang claim that their training is neutral with respect to the order of reasoning used by the subject. The training items can be answered successfully using

either first or second order reasoning, and therefore they should not affect the order of reasoning used by the subject. But would a second order reasoner really continue using second order reasoning for all 16 (or more) items, without discovering a much simpler (first order) heuristic with which he can perform just as well? A few first order heuristics that would results in correct predictions for all of Hedden and Zhang's training items are: "the opponent moves if cell C has a higher payoff than cell B, otherwise he stops", "the opponent moves if cell D has a higher payoff than cell B, otherwise he stops", and even "the opponent moves if cell B contains a payoff higher than 2, otherwise he stops". I fear that the training may encourage subjects to use one of these heuristics, and that the attempted transfer of such a heuristic to the diagnostic items may influence results.

I will use a modified training in which the items have only two decision points and three end points. At the first decision point the human player decides, at the second decision point the computer opponent decides. The training consists of 20 items. The first four items are 'familiarization'-items in which the subject is not asked to make predictions. In the remaining 16 items the subject must predict the opponent's action before making his move, just like in Hedden and Zhang's experiment. The payoffs used in the training are 1, 2, and 4. All 12 different combinations in which the human player starts with a payoff of 2 are included, and 8 games with a starting payoff of 1 or 4 are included as well (see Appendix A.1 for a list of these items). The last 6 games of the training are used to determine inclusion of the subject in the analysis of the remainder of the game. After completion of this training the subject is presented another four games to familiarize himself with the new, more complicated game board, since there is now an additional decision point. These familiarization games include one game starting with a payoff of 1, and three trivial games from Hedden and Zhang's original training. The subject is not asked to make predictions during this second familiarization. I believe that with my training set, the rather obvious change in the game between the training and testing phase makes transfer of simple heuristics impossible, while the familiarization phase with the 'real' game board is too short to allow the development of such heuristics.

4.2.3 Discussion

A problem with Hedden and Zhang's first order strategy

Hedden and Zhang specify very precisely to what predictions and actions the different strategies should lead. I agree that for a game of this 'depth' there is only one correct second order strategy. But I do not think that their zeroth and first order strategies are the only possible or viable strategies. I agree with Colman (2003) that Hedden and Zhang's characterization of zeroth order behaviour is problematic. A zeroth order player is a player who takes into account the state of the world and his own desires. But in Hedden and Zhang a zeroth order player looks only at the first two cells of the game and no further. If a player looks at his own payoffs at all four cells of the game, he would still

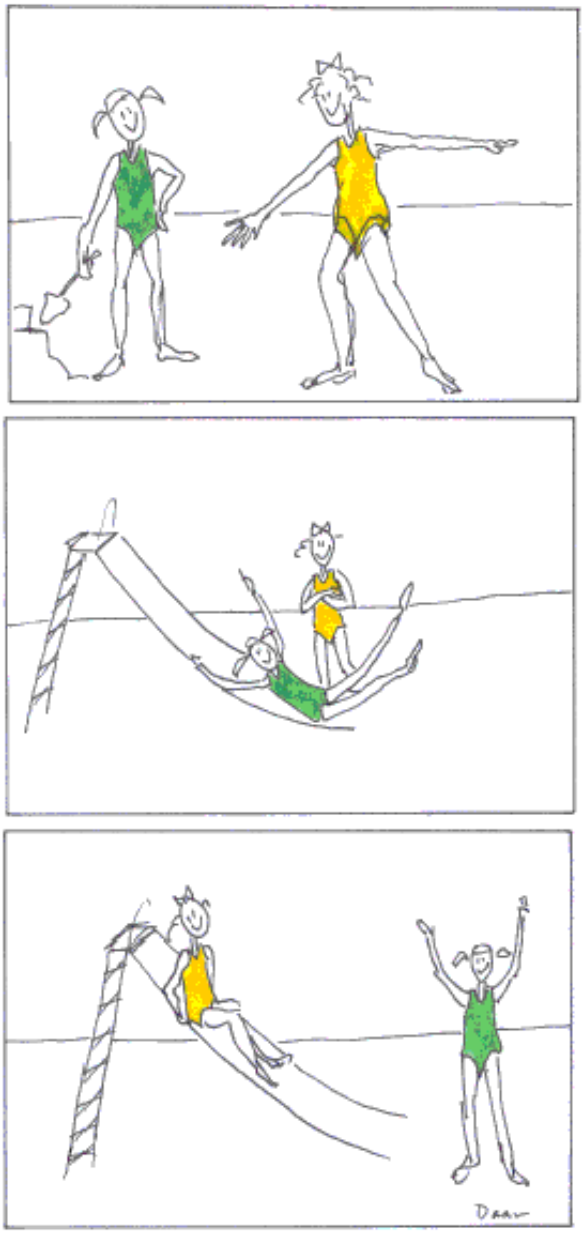
be zeroth order, since he still doesn't take his opponent's desires and payoffs into account. The zeroth order player would not know for certain in which cell the game would end if he moved, but he could maximize, minimize or average the uncertain payoffs to reach a decision, and it would still be zeroth order. These objections also apply to the (assumed) zeroth order behaviour of the opponent, leading to alternative first order strategies as well. In a matrix or ladder design, Hedden and Zhang's 'truncating' zeroth order strategy is certainly a possibility, but not the only possibility. In a tree design, truncating is more problematic. Why should the opponent compare cell B and C while ignoring D, if cell C and D are both equally far removed from the last decision point? A tree design would stimulate the opponent to find an alternative solution (such as averaging or maximizing cell C and D before the comparison with cell B), and it would stimulate a first order player to assume such a solution. But there is no reason why some players might not do the same thing in a matrix or ladder design - after all, the games are equivalent from a mathematical point of view.

What happens if we consider other first order strategies? A possible strategy would be: predict that the opponent moves towards his maximum. If a player would use this strategy, he would answer half of all diagnostic items correctly (i.e. he would give the same answer as someone using a second order strategy). Another consequence would be that someone using such a strategy would answer some of the control items incorrectly, even though someone using Hedden and Zhang's first order strategy would answer all control items correctly.

Unfortunately, I did not fully appreciate the problem of 'symmetry' between cell C and D and the plausibility of alternative first order strategies before the start of the experiment. With the current set of items, identical to those of Hedden and Zhang, it is not possible to draw strong conclusions about the use of such alternative strategies. Even different results between diagnostic items and control items are hard to interpret, because the control items are not balanced in the same way that the diagnostic items are.

Other commentary on Hedden and Zhang's game

Another point of critique voiced by Colman (2003) is that Hedden and Zhang prompt their players to make a prediction of the opponent's choice before they make their own choice. This prompting is likely to improve performance. It makes the game somewhat more like the 'explicit' false belief task than like the Keysar et al. (2003) experiment, in which the use of theory of mind by test subjects has to be their own 'spontaneous' idea. When comparing the results of the game to other experiments, one should keep in mind that performance might be worse had there been no prompting. However, I do not think this prompting invalidates the experiment, if one is careful in interpreting the results. My main interest is in obtaining results that allow me to make comparisons within individuals between tasks, and to make comparisons between adults and children; these comparisions will still be possible.



Twee meisjes gaan een dagje naar het strand. Kijk, hier zie je ze op het strand.

"Two girls are going to the beach for a day. Look, here you see them on the beach."

Als eerste gaan ze naar de glijbaan. Het meisje met het groene badpak gaat als eerste van de glijbaan af.

"First they go to the slide. The girl with the green bathing suit goes down the slide first."

Daarna gaat het meisje met het gele badpak van de glijbaan af.

"Then the girl with the yellow bathing suit goes down the slide."

Figure 4.5: The first story in the language test. Drawings by Danielle Koks.

Het meisje met het groene badpak gaat liever spelen. Kijk, daar staat een klimrek!

"The girl with the green bathing suit prefers to play. Look, there's a climbing frame!"

Het meisje met het groene badpak klimt op het klimrek.

"The girl with the green bathing suit climbs the climbing frame."

Daarna klimt het meisje met het gele badpak op het klimrek.

"After that the girl with the yellow bathing suit climbs the climbing frame."

Figure 4.6: The second story in the language test. Drawings by Danielle Koks.

Prompting can also be a problem between different tests. This is the reason that I will conduct the false belief task, which is most explicit in asking for ToM-reasoning, last.

4.3 Language test

For the language test I have chosen to use indefinite subjects. The materials from this test are taken from Vrieling (2006).

In this test the child listens to two short stories. In each story it is described how two different girls both perform a certain action. After each story the child hears a sentence and has to decide whether this sentence is correct. Each child hears one sentence with a sentence-initial subject ('canonical') and one sentence with an existential construction.

Children in condition A hear a canonical sentence first and an existential sentence second. Children in condition B hear an existential sentence first and a canonical sentence second.

The stories are shown in figures 4.5 and 4.6. They are the same for both conditions.

After listening to the first story, the child hears one of the following sentences (depending on the condition):

A Een meisje ging twee keer van de glijbaan af.
"A (particular) girl went down the slide twice."
canonical

B Er ging twee keer een meisje van de glijbaan af.
"Twice a girl went down the slide."
existential

After the second story the child hears one of these sentences:

A Er klom twee keer een meisje op het klimrek.
"Twice a girl climbed the climbing frame."
existential

B Een meisje klom twee keer op het klimrek.
"A (particular) girl climbed the climbing frame twice."
canonical

The child is asked whether the sentence is correct ("Is dat goed?") and the response is recorded. Subjects do not read the sentence, because they might think that they are supposed to check the spelling. If a subject does not answer immediately, the sentence is repeated once.

4.4 False belief task

In looking for a suitable second order false belief task, I found a Dutch ToM-test battery for clinical practice that includes a second order false belief test (Steerneman et al., 2003). However, according to the norms provided with the test battery, only half of all children up tot 10 years of age answer the second order false belief question correctly. Muris (Muris et al., 1999), reporting on the same group of children used for establishing the norms for the test battery, found that children succeeded at a translated version of Perner and Wimmer's (1985) John and Mary task from 7 years of age. Since these last results are a lot more consistent with results from American children, I must conclude that additional factors must have influenced the second order false belief task in this Dutch test battery.

I chose to use Sullivan et al.'s (1994) 'Birthday Puppy' story, because the creators argue convincingly that the story places less demands on information processing and memory than previous stories used in second order tests. I used the version of this story reported in Tager-Flushberg and Sullivan citeTager-Flusberg1994. I did however omit the second order ignorance question, and the memory aid is less explicit. Since my subjects are significantly older than Sullivan's, I did not think these questions would be necessary. The other questions, such as the first order ignorance question, remain, because they also function as control questions: children who fail any of these questions may not have understood the story correctly. Sullivan et al. also describe that they created a second story about a chocolate bar, but give little detail. I wrote a second order story about a chocolate bar, based primarily on the first order story with this theme by Hogrefe and Wimmer (1986). Both stories were accompanied by drawings by my hand.

In Figure 4.4 the drawings accompanying the chocolate bar story are presented. The text of the story is given below. The Dutch original of this story, and the other story used, can be found in Appendix B.

> John and Mary are brother and sister. Here they are in the living room. Then mother returns from shopping. Mother bought some chocolate. She gives the chocolate to John. Mary doesn't get any chocolate, because she has been naughty. John eats some of the chocolate and puts the remainder in the drawer. He doesn't give any of the chocolate to Mary. That makes Mary angry. Now John goes to help mother in the kitchen. He is helping with the dishes. Mary is alone in the living room. John is in the kitchen. Because she is angry with John, Mary hides the chocolate. She takes the chocolate out of the drawer and puts it in the toy chest.
>
> John is busy doing dishes. He throws the fruit leftovers in the rubbish bin in the garden. Through the window he sees the living room. He sees how Mary takes the chocolate out of the drawer, and puts it in the toy chest. Mary does *not* see John.
>
> *Reality control question:* Where is the chocolate now?

Figure 4.7: The drawings accompanying the chocolate bar story

1st order ignorance: Does John know that Mary has hidden the chocolate in the toy chest?
Linguistic control: Does Mary know that John saw her hide the chocolate?

John has finished the dishes. He is hungry. Now he wants to eat some of his chocolate. John enters the living room. He says: "Hmm, I would like some chocolate."

2nd order false belief: Where does Mary think that John will look for the chocolate?
Justification: Why does she think that?

The child's answers to each question are recorded. All adults received the story about the birthday puppy first. Since adults' performance on both stories was equally near-perfect, I did not realize that story order might be important. I started reversing the story order after administering the experiment at the first school, when I noticed a somewhat higher rate of error on the birthday puppy story than on the chocolate bar story. In the end, half of all children have the birthday puppy story first, and half heard the chocolate bar story first.

Chapter 5

Results

Each subject participated in three tests: the strategic game, the sentence comprehension test, and the second order false belief test (in that order). In the sentence comprehension test items were balanced so that half of the subjects received first an existential sentence and then a canonical sentence, the other half first a canonical sentence and then an existential sentence. For the child group, the order of the stories in the false belief test was also balanced. The tests were administered in one session that took about 30 minutes. The strategic game was played on a laptop computer with a separate mouse.

For none of the three tests differences between boys and girls or between the two schools that participated were found. In the next sections I will first describe the results for each test separately and then explore the relations between the tests.

5.1 Strategic game

5.1.1 Results from the training session

I have looked at the number of errors in the last six training items. I use six items because this results in a balanced mix. The last items were used so that the subject first has a chance to learn the game during the training. If the subject made an incorrect prediction, this was counted as a prediction error. If the subject made a correct prediction, but then chose an action that did not maximize his payoff, this was counted as a rationality error. Figure 5.1 shows the proportion of errors that were made during the last six training items. Adults answered 97% of the items correctly, children answered 71% correctly. In Figure 5.2 prediction errors and rationality errors are added together, and the number of errors per subject is shown.

The adults do a lot better than the children, with only one adult, but 18 of the 40 children (45%), making more than one mistake. Of the 18 children who made more than

Figure 5.1: Proportion of prediction errors, rationality errors, and correctly answered items for the last six items of the training.

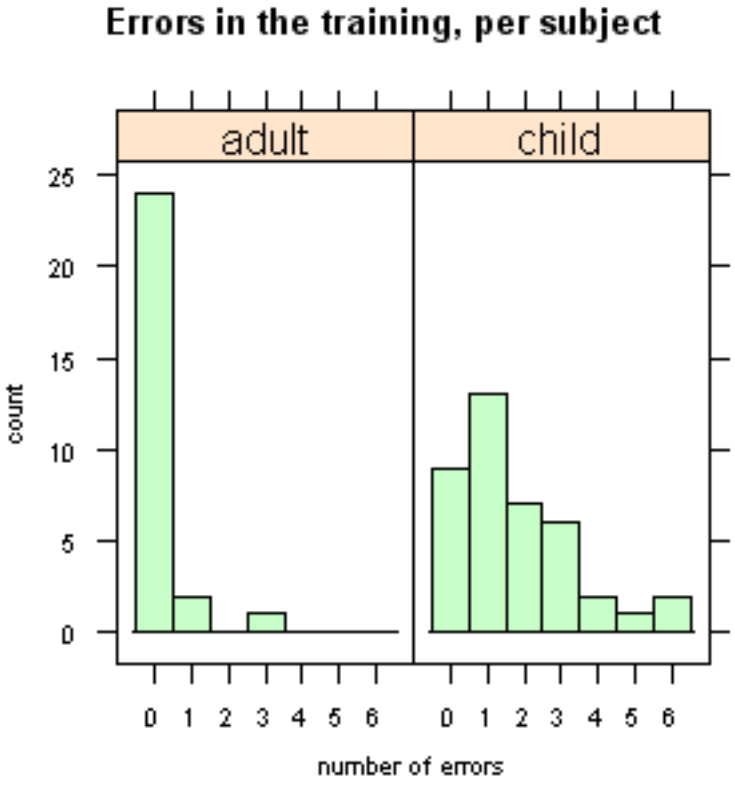


Figure 5.2: Histogram showing the number of errors per subject (prediction errors and rationality errors taken together) for the last six items of the training.

one mistake, 9 make at most one prediction error (see Appendix C for details). I think these children are capable of first order predictions, but have misunderstood the goals of the game. For now I can conclude that the majority of children are capable of making first order predictions (31 of 40 children have no more than one prediction error), but that these predictions are not always used to find the correct action.

5.1.2 Exclusions

I will exclude from the second order analysis anyone who made two or more errors in the last six training items.

Why exclude at all? The diagnostic items have only two possible responses, and are designed to distinguish between first and second order reasoning. Only if we know that somebody has mastered first order reasoning can we claim that second order answers in the test are due to second order reasoning. Mistakes in the training indicate an inability to correctly apply first order reasoning. The same reasoning that leads to mistakes in the training (whether this is the result of a misunderstanding of the goals of the game, the result of some kind of zeroth order reasoning, or the result of guesswork) may 'accidentally' result in a 'second order' response in the test in the absence of true second order reasoning. I chose to include subjects with at most one error in the training, because one error can be the result of a minor lapse of attention. But two errors out of six items (33%) is too many. Hedden and Zhang (2002) had a different exclusion criterium. They excluded subjects with at least two prediction errors or at least six rationality errors in the last 16 items of their training. For me, it's not possible to analyse that many 'last' items because I have fewer training items than Hedden and Zhang. Hedden and Zhang think that rationality errors are less serious than prediction errors. I do not agree. If a subject can not for himself choose rational actions in a first order situation, he can not predict the rational actions of his opponent in a second order situation. Therefore I have chosen to treat prediction errors and rationality errors the same, and exclude anyone with two or more errors.

By this exclusion criterium, 1 adult and 18 out of 40 children should be excluded. This difference between adults and children is highly significant ($\chi^2 = 13.531$, $p < 0.001$).

I have also looked at the player's decision at the last intersection of the diagnostic items and judged its 'correctness'. If the car reaches the last intersection, the subject chooses between the rewards at cell C and cell D. Because the opponent can not influence the outcome, the human player does not need to reason about the opponent's actions. Therefore, this intersection provides a simple choice between two alternative rewards that can be used to check whether a subject is employing appropriate goals for the game. If a subject is found to be using inappropriate goals, his predictions become difficult to interpret, because this subject will probably assume that his opponent uses the same inappropriate goals, or at least that his opponent assumes that he himself is using these inappropriate

goals. Thus, such a subject may make incorrect predictions despite using second order reasoning, or he may accidently make correct predictions without second order reasoning. Since the last intersection is never reached if the human players chooses to turn right at the first intersection, the numbers of times that the last intersection is reached differs from player to player and the absolute number of errors at the last intersection is not a good measurement. It was decided that a player should be excluded if more than 20% of the decisions at the last intersection were incorrect, with a minimum of three incorrect decisions. By this criterium 10 out of 40 children should be excluded, and none of the adults. Of the 10 children who should be excluded, 9 were already excluded based on the training, so the combined number of excluded children is 19 out of 40 (47.5%).

It is interesting to investigate the incorrect actions at junction C. The items at which more than one person made a mistake, were: 22, 32, 37, 52, 54, 61, 62, 63. I looked at the payoff structure for these items, and in all cases, the 'incorrect' action will actually maximize the *difference* in payoff between the player and the opponent. The 'incorrect' action leads to a lower payoff for the player than the 'right' action would, but the opponent suffers even more. If the goal of the game were to collect more marbles than the opponent, this action would be rational. Thus, these actions are consistent with competitive goals, as described in Section 4.2.2. My instruction and reward structure tried to emphasize that the goal of the game is to collect as many marbles as possible, regardless of how many the opponent collects. It seems that some children have ignored this instruction, or were unable to 'overrule' their inappropriate goals.

Appendix C provides more details on the 19 children and 1 adult who had to be excluded from second order analysis. The high number of rationality errors[1] and errors at the last intersection suggest to me that many exclusions were not the result of a lack of first order reasoning, but have to do with misunderstanding the mechanism or goals of the game. But, whether the main problem for a subject is a lack of first order reasoning or a misunderstanding of the goals of the game, the subject should be excluded from the second order analysis because, if included, his results could not be interpreted.

There is a significant correlation between the number of errors in the training and the proportion of incorrect actions at the last intersection (Pearson's correlation coefficient $R = 0.49$, $p < 0.001$)[2]. My main measurement in the next section will be the proportion of correct predictions on the diagnostic items. A high number of errors in the training correlates with a low score on this measure: $R = -0.36$, $p < 0.01$. The same is true for incorrect actions at the last intersection and correct predictions on the diagnostic items: $R = -0.43$, $p < 0.001$. I think these correlations support the validity of my exclusion criteria.

[1] Only correctly predicted items could be counted as rationality errors, so although the number of rationality errors is about equal to the number of prediction errors, their relative proportion is higher.

[2] An explanation of the statistical terminology used in this chapter can be found in many books on statistics. I recommend Palumbo (1977).

5.1.3 Second order predictions

Included in the analysis of the second order part of the game are 26 adults and 21 children.

The game had 32 diagnostic items. The number of correctly predicted items was calculated for each subject, so that each subject could be assigned a prediction score between 0 and 32. Prediction scores can be divided by 32 to obtain the proportion or percentage of correctly predicted items.

For the child group the mean prediction score was 18.29 (57.2%), the median was 17.00 (53.1%), and the standard deviation was 5.68.
For the adult group the mean prediction score was 24.15 (75.5%), the median was 25.5 (79.7%), and the standard deviation was 5.62.
A t-test was performed, which showed that the difference between the means was significant ($t = 3.54$, two-sided $p < 0.001$).

Figure 5.3 shows the distribution of prediction scores for each group. A curve representing the binomial distribution ($n = 32$, $\pi = 0.5$) that we would expect if subjects were guessing is also included in the figure. The child mean of 18.29 is significantly higher than the mean of 16 that we would expect if all subjects were guessing (one sample t-test, $t = 1.85$, one-sided $p = 0.04$), but only for a one-sided test[3]. The adult mean of 24.15 is of course also significantly higher than 16 (one sample t-test, $t = 7.39$, $p < 0.0001$). However, not all *individual scores* are above chance performance. How these scores should be interpreted is discussed in Section 6.1.1.

There is a bias towards 'go straight ahead' predictions. On average subjects predict that the opponent will go straight ahead for 59% of the diagnostic items, whereas this prediction is correct for only 50% of the diagnostic items.

5.1.4 Rationality

Most of the time subjects chose an action that is consistent with their prediction, but sometimes an incorrect action was chosen despite a correct prediction. As in the training phase, I call these actions 'rationality errors'. The proportion of rationality errors (as a percentage of all correct predictions, including control items) is 4.8% for adults, 15.5% for children. I also looked at the rationality of actions after an *incorrect* prediction[4]. For

[3]The one sample t-test only cares about the mean of the 'hypotheticial' distribution that the data is compared with, in this case 16. It estimates the standard deviation from the data. Since the actual standard deviation of our hypothetical binomial distribution ($n\pi(1-\pi) = 2.83$) is a lot lower than the standard deviation of the data (for both groups), this test probably *overestimates* the p-value.

[4]The rational action after an incorrect prediction is not automatically the opposite of the rational action after a correct prediction. The rationality of actions was computed as follows: if the player had predicted that the opponent would move to cell B, the player's payoff in cell B was compared with his payoff in cell A. If cell A had a higher payoff, then the rational action for the player was to turn right towards cell A, otherwise the rational action was to continue straight ahead. If the player had predicted

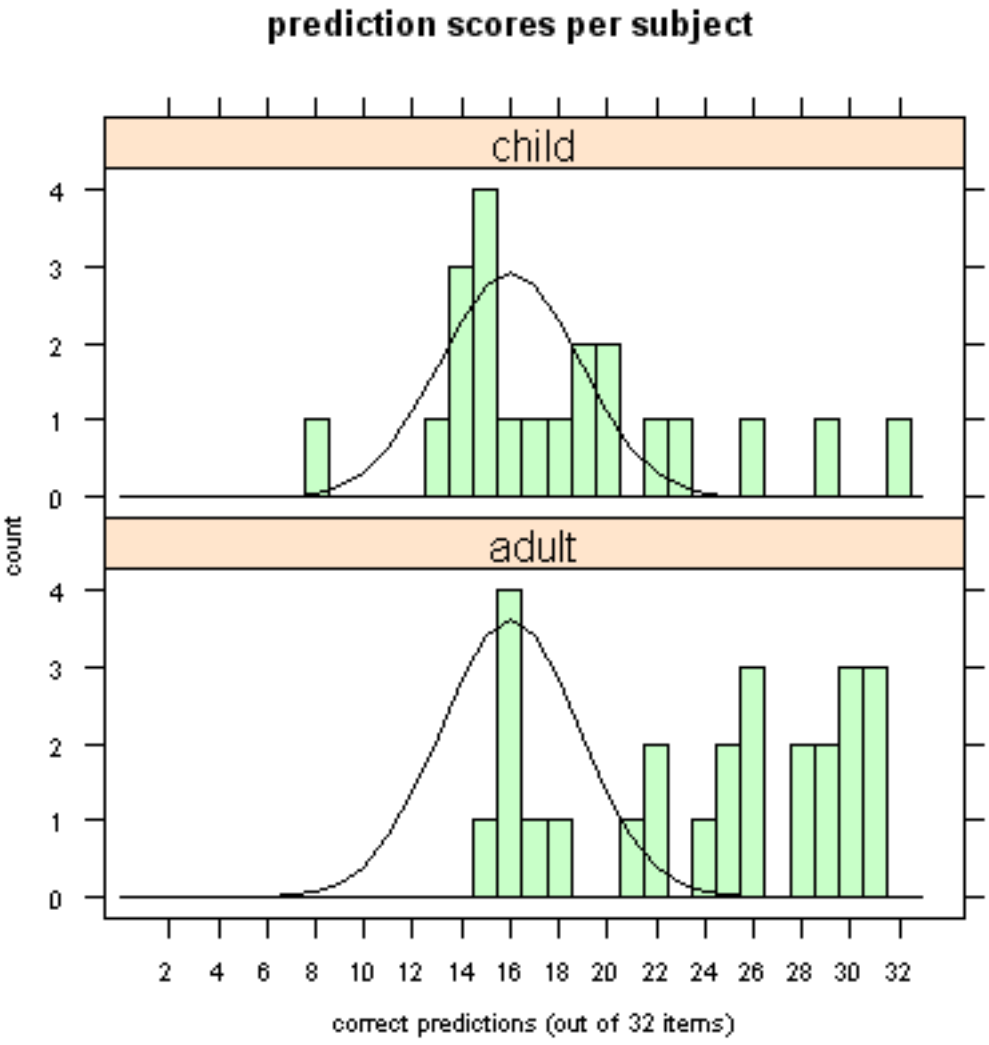


Figure 5.3: Histograms showing the prediction scores for each subject. The maximum obtainable prediction score was 32, because there were 32 diagnostic items. The black curve in the histogram represents the (binomial) distribution of scores that we would expect if subjects were guessing.

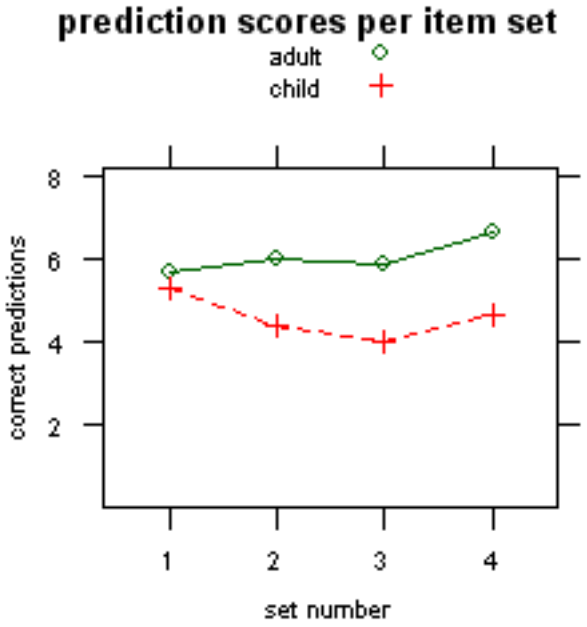


Figure 5.4: Average prediction scores per set. The upper line represents the adults, the lower line represents the children.

adults, 14.4% of actions was irrational given the incorrect prediction, for children this was 17.6%. The higher proportion of rationality errors for adults may mean that, rather than acting irrationally, some adults realized that their prediction was flawed and corrected themselves.

The rate of rationality errors for children in my experiment is comparable to the rate found by Hedden and Zhang for undergraduate students (for which I can't report a precise value because it was only reported in graphs, not in the text). My adult subjects however do a lot better than the subjects in Hedden and Zhang's experiment.

Whereas only 37% of all rational actions are a move to the right (see Figure 4.4 on page 36, when a rationality error was made, the action was a move to the right 51% of the time. Since moving to the right ends the game immediately and has a guaranteed outcome, this small bias to the right (not to be confused with the bias to *predict* that the *opponent* will go straight ahead) could be the result of risk-avoiding behaviour.

5.1.5 Learning effects

The 32 diagnostic items are divided in four balanced sets. For each set, the number of correct predictions can vary from 0 to 8. Figure 5.4 shows the average prediction score per set. A mixed effect regression analysis was carried out on the scores per set, with 'set' and 'group' as the dependent variables ('group' as a dummy, or indicator variable). Since there are multiple measurements per subject, the subject number was specified to be the random variable. The adults showed a small increase in correct prediction rate during the experiment ($\beta = 0.28$, $p = 0.0095$). But the children actually showed a decrease in prediction rate during the experiment ($\beta = -0.24$, $p = 0.056$). Although the increase for the adults is significant, the size of the effect is very small. Therefore, in the subsequent analysis of correlations between the strategic game and other tasks, I will use the overall prediction scores, not the scores per set. This has the additional advantage that it is 'safer' statistically, since the prediction score for the entire game, with 33 possible values from 0 to 32, is much closer to a continuous variable than the prediction score per set, which has only 9 possible values.

5.1.6 Control items

It was discovered only after finishing the design that the control items copied from Hedden and Zhang are not balanced in the same way as the diagnostic items are. For 16 of the diagnostic items the correct prediction is that the opponent will go straight ahead and for the other 16 it is that the opponent will turn right. But for only 2 out of 8 control items the correct prediction is 'go straight ahead'. In my experiment the proportion of correct predictions is lower for control items than for diagnostic items: 44.6% for children (diagnostic items: 57.1%) and 68.3% for adults (diagnostic items: 75.5%). The difference may be connected to the overall 'go straight ahead' prediction bias: 59% of all predictions for diagnostic items are 'go straight ahead' predictions, but only 50% of the items require such a prediction. For control items only 53% of all predictions are 'go straight ahead' predictions, but since such a prediction is correct for only 25% of the control items, the 'go straight ahead' bias may have contributed to the higher rate of error in control items compared to diagnostic items. This makes it difficult to compare control items and diagnostic items.

As argued in Section 4.2.1 a score on the control items significantly below 100% calls into doubt the first order strategy described by Hedden and Zhang. If the only first order strategy used by subjects would be the one described by Hedden and Zhang, the correct prediction rate for control items should be close to 100%. My results show that subjects are using other strategies than those proposed by Hedden and Zhang. The consequences

that the opponent would move straight ahead, the maximum of the player's payoffs in cell C and D was used instead of the player's payoff in cell B, and this number was compared with his payoff in cell A as before.

of this for the interpretation of prediction scores are discussed in Section 6.1.1.

5.2 Sentence comprehension

25 adult subjects were included in the sentence comprehension test, 12 in condition 1 and 13 in condition 2. All 40 children participated, 20 in condition A and 20 in condition B. Each subject heard one sentence with the subject in the canonical, sentence-initial position and one sentence with an existential construction. The sentences used can be found on page 43. The results are shown in Figure 5.5.

Adults always judge the existential sentences ("Er ging twee keer een meisje van de glijbaan af" or "Er klom twee keer een meisje op het klimrek") to be correct with regard to the story and picture given (see Section 4.3), which means that they assign a non-referential reading to the indefinite subject 'een meisje' (a girl). But only 16 out of 40 children thought the sentence was correct; the other 24 children assigned a referential reading to the subject. Of the 24 non-adultlike responses 9 were given in condition A (existential sentence last) and 15 were given in condition B (existential sentence first). A few children initially avoided a yes/no answer and gave a description of what happened in the story instead, but when explained that I wanted to know if I could use this particular sentence, they would give a yes/no answer. When children were asked why they thought the sentence was incorrect (I asked a few children in condition A, who heard the existential sentence last), they would explain that two different girls climbed. One girl thought that the verb in the existential sentence should be plural.

For the canonical sentences ("Een meisje ging twee keer van de glijbaan af" or "Een meisje klom twee keer op het klimrek"), children and adults show similar patterns of acceptance. 3 out of 25 adults and 6 out of 40 children consider the sentence correct, the others consider the sentence incorrect.

The difference between adults and children for the existential sentences is highly significant (χ^2=23.78, $p < 0.00001$).

5.3 False belief task

All subjects listened to two second order false belief stories. Since I expected nearly perfect performance on this task, I did not balance the order of the stories initially. One story was about a puppy that was to be a boy's birthday present. The other story was about a boy's chocolate bar that was hidden by his sister. The stories and the accompanying questions can be found in Section 4.4 and in Appendix B. When I found that the first story presented difficulties to some children, I started balancing the stories. Half of the

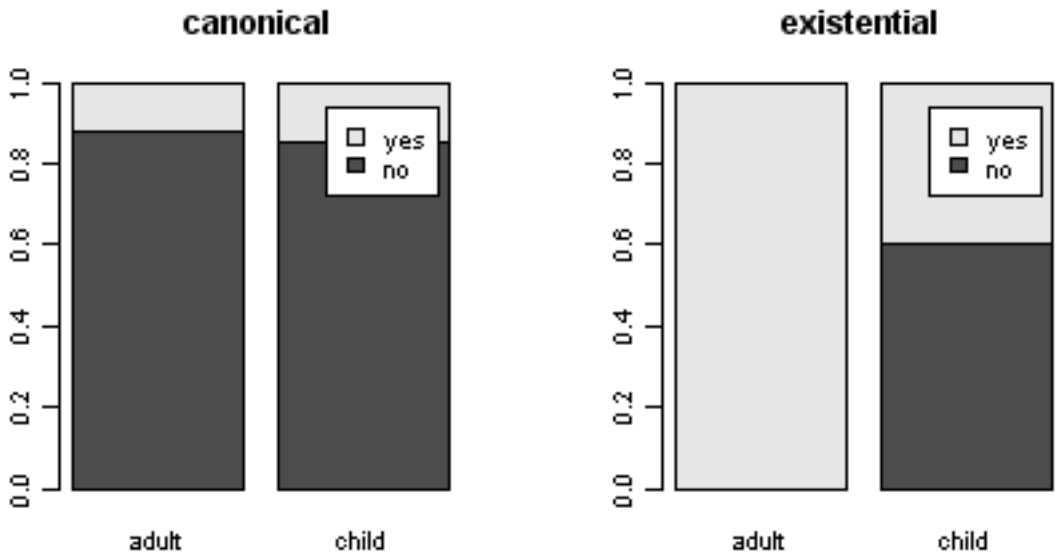


Figure 5.5: Correctness judgements for canonical sentences and existential sentences.

children heard the puppy story first, the other half heard the chocolate story first. All adults heard the puppy story first.

5.3.1 Adults

One adult gave an incorrect answer to the reality control question of the puppy story. The remainder of the story was skipped and the false belief question was not asked. Except for this one case, the second order false belief questions were answered correctly by all adults (53 cases). In one instance an incorrect justification was given to the puppy story.

The subjects sometimes used other words than those provided in the story. Alternatives for 'speelgoedkist' (toy chest) were 'speelgoeddoos' (toy box), 'doos', and 'kist'. Alternatives for 'la' (drawer) were 'ladenkastje' and 'kastje' (dresser, cabinet). Pointing to the right object in the picture was also counted as a correct response.

5.3.2 Children

In the child group one child answered the reality control question about the chocolate story incorrectly. One child answered the first order ignorance questions for both stories incorrectly, and three other children answered the first order ignorance question for the puppy story incorrectly. Two children answered the linguistic control question for the

	puppy story			chocolate story		
	n	**correct**	**justified correct**	**n**	**correct**	**justified correct**
children	36	72%	56%	36	92%	83%
adults	27	100%	97%	26	100%	100%

Table 5.1: Correct and justified correct responses to the second order false belief question for each story.

chocolate story incorrectly. All children with an incorrect answer to any of these questions were excluded from further analysis.

The puppy story was more difficult for the children than the chocolate story. 10 out of 36 children answered the second order false belief question about the puppy story incorrectly. Of the 26 children who answered correctly, 20 could give a correct justification.

Most incorrect answers to the false belief question were given with confidence and with a justification. The following conversation demonstrates such an incorrect answer. It seems that this child may not have understood the second order nature of the question:

Exp: Then grandma asks mother over the phone: "What does Peter think you bought him for his birthday?" What does mother answer to grandma?
Child: A puppy.
Exp: Why does mother say that?
Child: It's OK to tell grandma because *she* won't say anything.

Other justifications for incorrect answers were: "Because she told him what she bought", "Because it's grandma, not Peter.", "Because Peter wants a puppy".

The chocolate story was easier. The second order false belief question was answered correctly by 33 out of 36 children, 30 of which provided a correct justification. Of the three children who did not answer the false belief question correctly, only one had answered the false belief question for the puppy story correctly. The difference between the two stories was significant ($\chi^2 = 5.14$, $p < 0.05$).

Of the 13 incorrect responses to both stories, 10 were responses to the first story heard by that subject and only 3 were responses to the second story. This effect of story order is significant ($\chi^2 = 4.15$, $p < 0.05$). It is likely that children do better if they have already heard a similar story. It is also possible that children were influenced by the previous test and focusing on exact language and wording rather than on the contents of the story.

The responses for both stories are tallied in table 5.1. The children's performance on the puppy story is consistent with Perner and Wimmer (1985) and on the chocolate story it is somewhat better. Although subsequent research with simplified stories (on which my stories were based) finds higher performance than Perner and Wimmer for young children, this research does not investigate children in the age group tested by me, so no direct comparison can be made.

5.3.3 Justifications

By far the most common justification for the correct answer 'a soccer ball' for the puppy story was "Because that's what mother said to Peter." Other justifications that occurred several times were: "Because mother doesn't know that Peter saw the puppy," "Because she wanted to surprise him," and "Because mother doesn't know that Peter knows that she bought a puppy." Only that last justification is explicitly second order.

The most common justification for the correct answer 'the drawer' for the chocolate story is second order: "Because Mary doesn't know that he saw that she hid the chocolate." The other common justification was zeroth order: "Because John put it there (in the drawer)."

5.3.4 Difference between the stories

As described above there is a significant difference between the stories. The chocolate story is easier than the puppy story. Although for the puppy story the difference between adults and children is significant ($\chi^2 = 8.61$, $p < 0.01$), for the chocolate story it is not. Compared with the original birthday puppy story by Sullivan et al. (1994) my story is slightly shorter. It lacks the second order ignorance question, although a memory aid is still incorporated in the story. However, the chocolate story has the same structure, so this can not explain the lower results. Maybe the puppy story is more difficult because it features more dialogue than the chocolate story. This dialogue is harder to depict in the drawings and therefore harder for the child to remember.

My interpretation of these results is that something is wrong with the puppy story, and that the children's performance on the chocolate story is a better measurement of their ToM-reasoning ability in a false belief task. Therefore in the analysis of correlations between tasks I will use the chocolate story only.

5.4 Correlations

5.4.1 The strategic game and sentence comprehension

Since the number of 'deviant' responses to the canonical sentences is very low and the proportion is similar for both adults and children, I do not think it fruitful to look for correlations with this response. I will look at the response to the existential sentence only. Since adults have a uniform 'yes' response to this sentence, and we already know that adults score higher on the strategic game than children, I will look only for correlations within the child data. If adults and children were included in the same analysis and a correlation was found, it might be spurious.

There is no significant difference in prediction scores for the strategic game between those children who considered the existential sentence incorrect and those who considered it correct. The histograms in Figure 5.6 show the prediction score on the strategic game for children. The upper histogram gives the scores for those children who, like adults, considered the existential sentence correct, the lower histogram for those who, unlike adults, considered the existential sentence incorrect. The mean prediction score for the adultlike group is 16, for the non-adultlike group it is 19.6. The prediction scores for both groups were compared using a two sample t-test with pooled variance, and it was found that $t = 1.49$, $p = 0.15$[5]. A test specifically suited for small, non-normal samples might find an even higher p-value. Furthermore, the difference between the groups is not in the direction that I predicted or can explain: children with an adultlike response to the sentence task perform *worse* on the strategic game than children with a non-adultlike response. Thus, despite sufficient variation in the data for both tasks, I can not prove a relation between a child's score on the sentence comprehension task and a child's score on the strategic game.

Because of the difficulty of interpreting and comparing the lower prediction scores (those around 50%), as explained in Section 6.1.1, I also divided the children into 'those who score significantly higher than chance on the strategic game' and 'those who do not'. A score of 22 (out of 32) or more correct predictions was considered higher than chance. Please note that there is no natural divide in the data around this score; it is based on binomial probabilities (for individual measurements, $p(x < 22) = .025$, while $p(x < 21) = 0.055$). Fisher's exact test on this factor, crossed with the response to the existential sentence, gives a p-value of 0.11[6]. Again, the effect is not significant.

5.4.2 The strategic game and the false belief task

Three children answered the second order false belief question about the chocolate story incorrectly. Of these three children only one was included in the strategic game analysis, and had a prediction score of 14. This prediction score is not evidence of second order reasoning (see Section 6.1.1). None of the three additional children who could not justify their correct answer were included in the strategic game analysis either. These results are consistent with the assumption that passing a second order false belief task is a necessary condition for applying second order reasoning to other tasks such as the strategic game,

[5]Figure 5.6 shows that the adultlike group had very little variance in their prediction scores. I think this low variance is a coincidence that was possible because of the small sample size, and that it would be very unreliable to use this variance in the t-test. Therefore I conducted the t-test under the assumption of equal variances, so that the test uses *one* variance based on all the data from both samples. If the t-test is conducted with the assumption of unequal variances, the outcome of the test is that $t = 1.80$ and $p = 0.09$

[6]Fisher's exact test is an alternative for the χ^2-test when the sample size is not large enough for the χ^2-test to be reliable. The sample size for this test was small, because almost half of the children were excluded from the second order analysis of the strategic game.

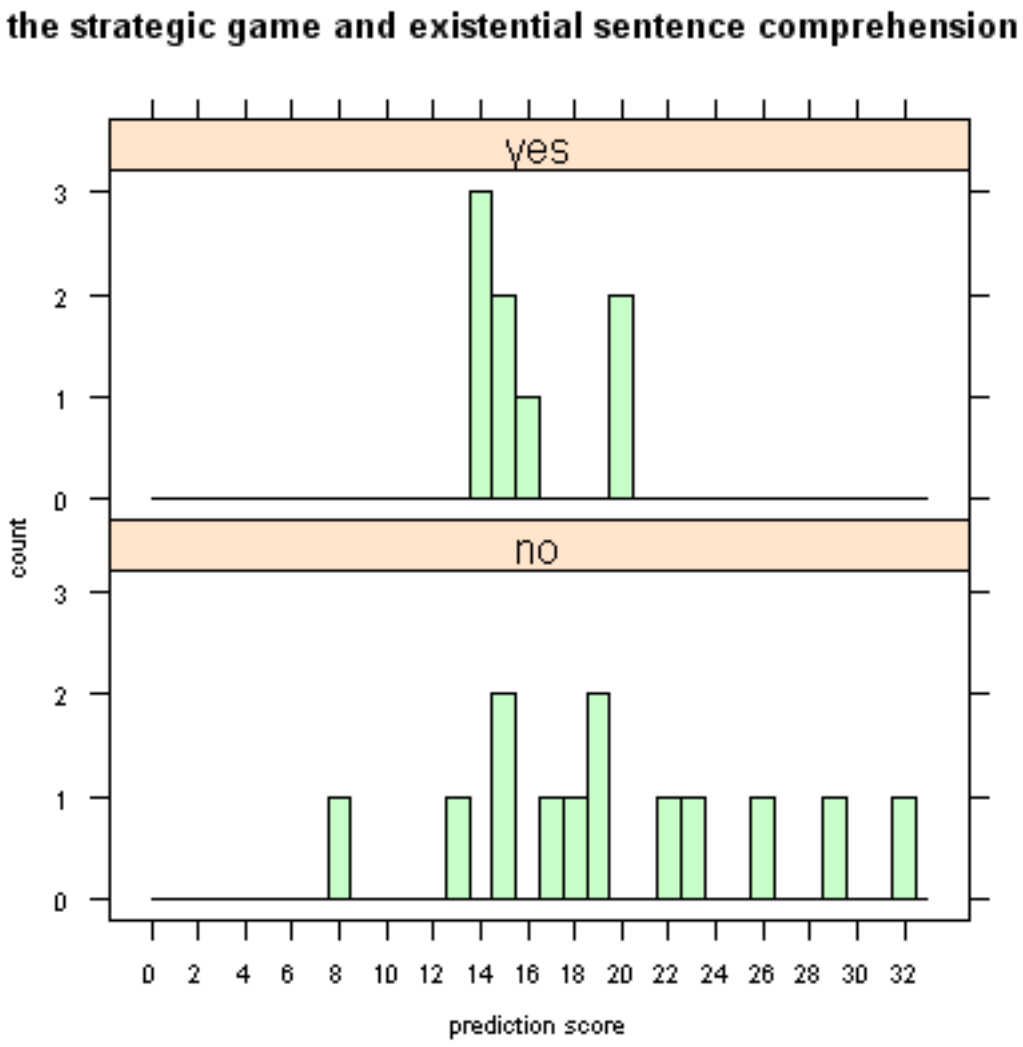


Figure 5.6: The upper histogram shows the distribution of prediction scores on the strategic game for children with the adultlike 'yes'-response to the existential sentence. The lower histogram shows the prediction scores for the children with the non-adultlike 'no'-response.

but this assumption cannot be proven because there is insufficient (variation in the) data. Since for both adults and children performance on the game is far from perfect, it can be concluded that success at the false belief task is not *sufficient* for success at the strategic game.

5.4.3 Sentence comprehension and the false belief task

Of the three children who answered the second order false belief question about the chocolate story incorrectly, one had an adultlike interpretation of the existential sentence and two had an non-adultlike interpretation. This distribution of responses is similar to the rest of the child data (40% non-adultlike interpretation), but no conclusions can be drawn from such small numbers.

5.5 Summary of the results

At the end of the training phase of the strategic game the majority (77.5%) of the children and all adults are capable of making correct first order predictions.

Adults and children who are capable of applying first order reasoning in the training phase of the strategic game exhibit second order reasoning in the testing phase of the strategic game. Adults make more correct second order predictions (75.5%) than children (57.2%). In both groups performance is far from perfect. There are considerable differences between individuals.

Both adults and children reliably use their predictions in the strategic game to guide their actions.

There is no learning effect or strategy change during the game. People who use second order reasoning, do so from the start.

The majority (60%) of the children interpret existential sentences different than adults do: they interpret the indefinite subject as referential.

Almost all subjects succeed at a second order false belief task. Success at a false belief task may be necessary but is not sufficient for success at a strategic game or at a comprehension task of existential sentences. Since adults do better on the strategic game than children, applied ToM-reasoning must continue to develop after 'passing' a false belief task.

Despite sufficient variation in the data for the sentence comprehension task and the strategic game, there seems to be no relation between between a child's score on these tasks.

Chapter 6

Discussion

6.1 Comparison with Hedden and Zhang

6.1.1 Interpretation of prediction scores

In Hedden and Zhang's game (2002), prediction scores rise from an inital low value around 0.2 to about 0.6/0.7 towards the end of the test session. Hedden and Zhang's interpretation of these scores is that subjects start with a default first order strategy and gradually adopt a second order strategy. In between there is a point where subjects use first order reasoning about half of the time and second order reasoning the other half of the time. Hedden and Zhang do not report results on the control items, but my own results on control items indicate that Hedden and Zhang's interpretation is problematic: subjects make many mistakes on control items, which means they must be using other strategies than the first and second order strategy proposed by Hedden and Zhang.

What can be said about some of my own subjects who score around 50% *throughout the game*? Can I be sure that these subjects are using first order reasoning half of the time and second order reasoning the other half of the time? I made sure that subjects who could not apply first order reasoning, or who acted irrationally, were excluded from the second order analysis. Therefore, it is unlikely that any of the included subjects were guessing. But it is still possible that they are using an alternative first order strategy that works about half of the time (I argued that such strategies exist in Section 4.2.3).

I am reluctant to claim that my subjects use a second order strategy $x\%$ of the time, and a first order strategy $100-x\%$ of the time, where x is the subject's prediction score. I still believe that prediction scores higher than 50% can not be achieved without using second order reasoning at least some of the time. Although I don't know what to claim about subjects with prediction scores around 50%, I think it is safe to claim that many subjects *do* use second order reasoning (because their prediction scores are significantly higher than 50%) and that adults perform better than children. At the upper half of the scale it can

also be said that the higher a subject's score, the more he has used second order reasoning. For scores around or below 50% this need not be true, and the differences between these scores need not indicate a difference in ToM-reasoning ability. This was the reason that, when comparing results on the game to the sentence comprehension test, I divided my subjects in 'high' and 'low' scorers and performed χ^2-tests on this dichotomous data in addition to performing the t-test on the prediction scores.

6.1.2 Learning during the game

In Hedden and Zhang's experiment correct predictions started at a low rate of 20% and increased markedly during the experiment. In my experiment correct predictions start at a much higher rate and change little during the experiment. I can think of two possible explanations for this.

The first explanation is that my game uses a different and more concrete presentation. I intentionally made my game 'easier' so that it could be played by children. The changes I made may have benefited the adult subjects as well. Since my presentation is mathematically equivalent to that of Hedden and Zhang, the difference shouldn't be very important once the subject is thoroughly familiar with the game. But especially at the start of the game (where the difference between my results and Hedden and Zhang's is greatest) a better presentation may increase performance.

The other possible explanation is that the difference in results is caused by a difference in training items between Hedden and Zhang's game and mine. The initial reason for creating my own training items was simply to make the training shorter and fit the entire experiment within the attention span of a 9-year-old child. But once I had designed training items with only two intersections and three cells I started to like them for different reasons. I explained in 4.2.2 that I think inappropriate transfer from Hedden and Zhang's training items to the test session is possible. This inappropriate transfer could explain why Hedden and Zhang's subjects start at a low rate of second order predictions. They start the test session using the same strategy that served them well during the training. The increase in prediction scores during the experiment merely represents that they are 'unlearning' this strategy.

Of course it is possible that both of these explanations contribute to the difference in results between Hedden and Zhang's and my own research. A test of one or both of these explanations would be an interesting option for future research. This could be accomplished by repeating the game experiment in a *between subjects* design, using two different game presentations, or two different sets of training items.

The result that people who use second order reasoning do so from the start of the game was also found in an experiment by Mol et al. (2005). They used an experiment based on the game Mastermind. They found, to their surprise, that people did not learn new strategies during the game. Subjects who used second order reasoning did so from the

start of the game. They also identified a large group of subjects who used first order reasoning only, and these subjects were unable to switch to second order reasoning.

6.2 Competitive goals in the strategic game

In section 4.2.2 I explained that the goal of the strategic game, to obtain as many marbles as possible, is an egoistic goal, and that this is different from the competitive goals often encountered in board games or computer games. I found that many of my child subjects used inappropriate competitive goals. I was able to detect and exclude children who did so, but it is regrettable that so many children had to be excluded. Although adults can choose to use the appropriate egoistic goals if they are explicitly instructed to do so, this is not the case for all children. I do not think it is possible to adapt the game in such a way that competitive and egoistic goals always lead to the same actions; not without linking the player's payoff and the computer's payoff in a way similar to zero-sum games, in which case the necessity to reason about the opponent's payoffs would be lost. In future research it should be remembered that the potential conflict between competitive and egoistic goals needs to be investigated when designing experiments using games, especially if child subjects are used.

6.3 What makes applied ToM-tasks so hard

With the exception of a few subjects who obtained near-perfect prediction scores, performance on the strategic game was far from perfect for both adults and children. It would be interesting to investigate why most people do not always use their second order reasoning skills in applied tasks. Does the game require additional skills that are not needed in a false belief task, and if so, what kind of skills? Does a game situation encourage people to guess? Will people do better if the rewards are increased? Can subjects be trained to do better? All these questions seem to me interesting questions for future research. Our Theory of Mind is not complete if we do not know why and when people use their theory of mind.

6.4 Does the sentence comprehension task involve ToM?

My research was unable to prove a relation between the sentence comprehension task and the strategic game, despite sufficient variation in the data for each of these tasks. Therefore, I can't give any answers to this question on the basis of my experimental research and the question remains open.

6.5 Interpretation of canonical sentences

I claimed in Section 4.3 that in Dutch, indefinite subjects in canonical sentences receive a referential reading. A small minority of my subjects, both adults and children, seem to disagree (Section 5.2).

Apparently a minority of subjects consider a non-referential interpretation possible. This is a bit surprising but not unexplainable: in English a non-referential reading for canonical sentences ("A girl went down the slide twice.") is quite acceptable, perhaps because there is no other way (such as the existential construction) to easily express a non-referential reading in English. It is not the minority, non-referential reading of the canonical sentence that needs an explanation, it is the impossibility of such a reading to the majority of most Dutch speakers that needs an explanation. I would speculate that adult speakers have blocked the non-referential meaning, precisely because an alternative way to express this meaning exists in Dutch. I do not think children necessarily arrive at this referential reading through the same process as adults: it may be that they simply do not think about the non-referential reading for subjects in the first place.

Chapter 7

Conclusion

The aim of this thesis was to study the development and application of second order reasoning in the context of strategic games and in reasoning about speaker's alternatives. In the literature I found that children succeed at second order reasoning in a false belief task at age 6 or 7, but that even adults often have far from perfect performance on applied reasoning tasks. Therefore, my experimental research used both a second order false belief task and an applied, strategic game to measure the ability for applied second order reasoning. To investigate the possible link with language, I used a test on the comprehension of sentences with indefinite subjects. For correct interpretation of these sentences it is necessary that the listener considers the speaker's alternatives. My research compared adults with children aged 8-10 years.

I will now answer the questions formulated in Chapter 3.

> Is performance on a false belief task related to performance on a strategic game?

Almost all children succeeded at a second order false belief task. Those who did not had been excluded from or performed poorly on the strategic game. Given these results, it is very well possible that success at a second order false belief task is a necessary condition for and must precede second order reasoning in the context of a strategic game. However the number of children who failed the false belief task is too small to be certain of this.

Since performance on the strategic game is far from perfect for both adults and children, success at a false belief task is not sufficient for success at a strategic game.

> Is performance on a sentence comprehension task in which correct comprehension requires reasoning about the speaker's alternatives related to performance on other ToM-tasks (the false belief task or the strategic game)?

No relation between the sentence comprehension task and the strategic game was found, despite sufficient variation in the data for both tasks. No relation was found between

the sentence comprehension task and the false belief task, either, but performance on the false belief task was too uniform to find a relation if one existed.

What are the differences between adults and children for all three tasks?

On average adults do better on the strategic game than children.
Whereas all adults interpret the indefinite subject of an existential sentences non-referentially, 60% of 8-10 year old children still have a non-adultlike, referential interpretation.
For the second order false belief task the difference between adults and children is too small to be significant.

How does second order reasoning develop and how is it applied to strategic games and reasoning about speaker's alternatives?

Although children succeed at a second order false belief task at about 6/7 years of age, applied second order reasoning is an advanced skill that continues to develop after this age. Second order reasoning is applied to some degree to the strategic game, even by children aged 8-10, but performance is far from perfect. No relation between second order reasoning and reasoning about speaker's alternatives was found.

Chapter 8

Summary

Many everyday reasoning tasks require a person to reason about the knowledge, beliefs, intentions, and goals of other people. The approach known as Theory of Mind (ToM) assumes that the capacity for this kind of reasoning depends on a mental model of the knowledge and intentions of other people: a '*theory of mind*'. ToM-reasoning can be classified by its order of reasoning. A person capable of first order reasoning can reason about another person's beliefs, provided that these beliefs concern the state of the world: "(I know that) Mary thinks the ball is in the bag.". Someone capable of second order reasoning can also take into account another person's beliefs about others, including himself: "(I know that) Mary thinks that John thinks the ball is still in the cupboard."

In this research the development of applied second order reasoning was studied through experimental research on adults and children. In previous experiments with adult subjects, large performance differences between the standard false belief task and more applied tasks of ToM-reasoning were found. Therefore, this research used both a second order false belief task and a strategic game by Hedden and Zhang's (2002) to measure ToM-reasoning. In the strategic game a subject needed to apply second order ToM-reasoning to (correctly) predict his opponent's actions and maximize his own score. In the domain of language, reasoning about speaker's alternatives resembles second order ToM-reasoning. A sentence comprehension task on indefinite subjects was used to investigate the possible link between ToM-reasoning and language.

40 children (age: 8-10 years) and 27 adults participated in each of the three tasks described above. The analysis examined the differences between adults and children for each task, and it investigated whether an individual's performance on one of the tasks predicted his performance on the other tasks.

The second order false belief task showed no differences between adults and children, because in both groups nearly all individuals succeeded at the task. The strategic game showed large differences between the groups. The majority of the children (77.5%) and all adults were capable of making correct first order predictions at the end of the training

phase. During the testing phase, children made correct second order predictions 57.2% of the time, while adults made correct second order predictions 75.5% of the time. There are considerable differences between individuals, and for most individuals in both groups, performance is far from perfect. My results differ from Hedden and Zhang's, in that my subjects exhibit no learning during the testing phase. My subjects' performance *throughout* the testing phase is comparable to the performance of Hedden and Zhang's subjects *at the end* of the testing phase. This difference may be because I adapted the visual presentation of the game, or it may be because I changed the training items, which makes a transfer of (inappropriate) heuristics from training to testing phase more difficult. The sentence comprehension task also showed differences between adults and children: the majority (60%) of the children interpret indefinite subjects in an existential construction different than adults do.

There was no relation between a child's score on the sentence comprehension task and the strategic game. Because very few people failed the false belief task, it is not possible to conclude whether or not success at the false belief task is necessary for success at the other tasks. Since most individuals did not demonstrate perfect or near-perfect performance on the strategic game despite success at the false belief task, it can be concluded that success at the second order false belief task is not sufficient for success at the strategic game. A person who has mastered second order reasoning in the context of a false belief task does not necessarily apply these reasoning skills to other tasks. Adults apply second order reasoning more often than children. Although children succeed at a second order false belief task at about 6/7 years of age, applied second order reasoning is an advanced skill that continues to develop after this age.

Bibliography

Abelson, H., Dybvig, R. K., Haynes, C. T., Rozas, G. J., Iv, N. I. A., Friedman, D. P., Kohlbecker, E., Jr., G. L. S., Bartley, D. H., Halstead, R., Oxley, D., Sussman, G. J., Brooks, G., Hanson, C., Pitman, K. M., and Wand, M. (1998). Revised (4) report on the algorithmic language scheme. *Higher-Order and Symbolic Computation*, 11:7–105.

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

Blutner, R., De Hoop, H., and Hendriks, P. (2006). *Optimal Communication*. CSLI Publications, Stanford, CA (USA).

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

De Hoop, H. and Krämer, I. (2006). Children's optimal interpretations of indefinite subjects and object. *Language Acquisition*, 13:103–123.

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

Feeney, A., Scrafton, S., Duckworth, A., and Handley, S. (2004). The story of some: Everyday pragmatic inference by children and adults. *Canadian Journal of Experimental Psychology*, 58:121–132.

Flavell, J. and Miller, S. (2002). *Cognitive Development*. Prentice Hall, Englewood Cliffs, NJ (USA).

Hedden, T. and Zhang, J. (2002). What do you think I think you think? strategic reasoning in matrix games. *Cognition*, 85:1–36.

Hendriks, P. and Spenader, J. (2004). A bidirectional explanation of the pronoun interpretation problem. In Schlenker, P. and Keenan, E., editors, *Proceedings of the ESSLLI '04 Workshop on Semantic Approaches to Binding Theory*, Nancy (France).

Hendriks, P. and Spenader, J. (to appear). When production precedes comprehension: An optimization approach to the acquisition of pronouns. *Language Acquisition*.

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

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

Mol, L., Taatgen, N., Verbrugge, R., and Hendriks, P. (2005). Reflective cognition as a secondary task. In Bara, B., Barsalou, L., and Bucciarelli, M., editors, *Proceedings of the 27th Annual Meeting of the Cognitive Science Society*, pages 1525–1530, Mahwah,

NJ (USA). Erlbaum.

Muris, P., Steerneman, P., Meesters, C., Merckelbach, H., Horselenberg, R., van den Hogen, T., and van Dongen, L. (1999). The tom test: A new instrument for assessing theory of mind in normal children and children with pervasive developmental disorders. *Journal of Autism and Developmental Disorders*, 29:67–89.

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

Noveck, I. (2001). When children are more logical than adults: experimental investigations of scalar implicature. *Cognition*, 78:165–188.

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

Palumbo, D. J. (1977). *Statistics in political and behavioral science*. Columbia University Press, Guildford, NY (USA).

Papafragou, A. and Musolino, J. (2003). Scalar implicatures: Experiments at the semantics-pragmatics interface. *Cognition*, 86:253–282.

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

Perner, J. (1979). Young children's preoccupation with their own payoffs in strategic analysis of 2 x 2 games. *Developmental Psychology*, 15:204–213.

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

Perner, J. and 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.

Premack, D. and Woodruff, G. (1978). Does the chimpanzee have a theory of mind. *Behavioral and Brain Sciences*, 1:515.

Russell, J., Jarrold, C., Sharpe, S., and Tidswell, T. (1991). The 'windows' task as a measure of strategic deception in preschoolers and autistic subjects. *British journal of developmental psychology*, 9:331–350.

Russell, S. J. and Norvig, P. (1995). *Artificial Intelligence: A Modern Approach*. Prentice-Hall, Englewood Cliffs, NJ (USA).

Samuels, M. C., Brooks, P. J., and Frye, D. (1996). Strategic game playing in children through the windows task. *British journal of developmental psychology*, 14:159–172.

Steerneman, P., Meesters, C., and Muris, P. (2003). *TOM - test*. Garant Uitgevers, Antwerpen (Belgium).

Sullivan, K., Zaitchik, D., and Tager-Flusberg, H. (1994). Preschoolers can attribute second-order beliefs. *Developmental Psychology*, 30:395–402.

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

Tomasello, M., Kruger, A., and Ratner, H. (1993). Cultural learning. *Behavioral and Brain Sciences*, 16:495–552.

Van Rooij, R. (2004). Signalling games select Horn strategies. *Linguistics and Philosophy*, 27:493–527.

Vrieling, P. (2006). Een ezel stoot zich geen twee keer aan dezelfde steen: Dutch children's

interpretations of indefinite subject NPs. Master's thesis, Utrecht University.
Wimmer, H. and Perner, J. (1983). Beliefs about beliefs: representation and constraining function of wrong beliefs in young children's understanding of deception. *Cognition*, 13:103–128.

Appendix A

Strategic game experiment

A.1 Items

The items were administered in the same order for all subjects.

A.1.1 Training session

item	**player payoff**			**opponent payoff**			**correct pred.**	**correct act.**
	A	B	C	A	B	C	(2^{nd} junction)	(1^{st} junction)
1	1	2	4	2	1	4	move	move
2	4	1	2	1	4	2	right	right
3	1	2	4	1	4	2	move	move
4	2	4	1	2	1	4	move	right
5	2	1	4	2	1	4	move	move
6	4	2	1	2	4	1	right	right
7	2	1	4	1	4	2	right	right
8	2	4	1	2	4	1	right	move
9	1	4	2	2	4	1	right	move
10	2	1	4	4	2	1	right	right
11	2	1	4	4	1	2	move	move
12	4	1	2	4	1	2	move	right
13	2	4	1	4	2	1	right	move
14	2	4	1	1	4	2	right	move
15	2	4	1	4	1	2	move	right
16	1	4	2	2	4	1	right	move
17	2	1	4	1	2	4	move	move
18	4	2	1	2	1	4	move	right
19	2	4	1	1	4	2	right	move
20	2	1	4	2	4	1	right	right

For items 1-4 no predictions are asked.

A.1.2 Test session

item	player payoff				opponent payoff				correct pred.	correct act.	H&Z
	A	B	C	D	A	B	C	D	(2^{nd} junction)	(1^{st} junction)	
21	1	2	4	7	1	2	4	7	move	move	
22	4	7	2	1	2	1	7	4	move	right	
23	4	7	2	1	4	7	1	2	right	move	
24	2	1	4	7	7	4	2	1	right	right	
25	4	2	1	7	7	2	1	4	move	move	1;4
26	2	1	4	7	1	4	7	2	right	right	2;2
27	2	4	1	7	2	4	7	1	right	move	2;3
28*	2	1	7	4	7	2	1	4	right	right	2;4
29	4	1	2	7	1	4	7	2	right	right	1;5
30	2	7	1	4	7	2	1	4	move	move	2;1
31	4	7	1	2	2	4	7	1	right	move	1;1
32	4	7	1	2	4	2	1	7	move	right	1;2
33*	4	7	2	1	4	2	1	7	right	move	1;3
34	2	1	4	7	4	2	1	7	move	move	2;5
35	4	7	1	2	7	2	4	1	right	move	1;6
36*	2	1	7	4	2	4	1	7	right	right	2;9
37	4	7	1	2	2	4	1	7	move	right	1;10
38*	4	1	7	2	7	2	4	1	move	move	1;9
39	2	1	4	7	1	4	2	7	move	move	2;10
40	2	4	1	7	1	4	7	2	right	move	2;6
41	4	1	2	7	4	2	7	1	right	right	1;8
42	2	4	1	7	7	2	1	4	move	move	2;8
43	4	2	1	7	1	4	2	7	move	move	1;7
44	2	1	4	7	2	4	7	1	right	right	2;7
45	4	1	2	7	2	4	1	7	move	move	1;12
46	2	1	4	7	7	2	4	1	right	right	2;11
47*	4	2	7	1	1	4	2	7	right	right	1;13
48	4	7	1	2	4	2	7	1	right	move	1;14
49*	2	4	7	1	1	4	2	7	right	move	2;13
50	4	2	1	7	4	2	7	1	right	right	1;11
51	2	1	4	7	7	2	1	4	move	move	2;14
52	4	7	1	2	1	4	2	7	move	right	1;15
53	2	7	1	4	4	2	7	1	right	move	2;12
54	2	7	1	4	2	4	1	7	move	move	2;15
55	4	1	2	7	7	2	1	4	move	move	1;16
56	4	7	1	2	1	4	7	2	right	move	1;20
57	2	1	4	7	4	2	7	1	right	right	2;16
58	2	1	4	7	2	4	1	7	move	move	2;17
59	4	2	1	7	2	4	7	1	right	right	1;17
60*	2	1	7	4	4	2	1	7	right	right	2;18
61	4	7	1	2	7	2	1	4	move	right	1;18
62*	4	7	2	1	1	4	7	2	move	right	1;19
63	2	7	1	4	1	4	2	7	move	move	2;19
64	2	4	1	7	7	2	4	1	right	move	2;20

For items 25-64, the last column (labeled "H&Z") lists the block and item number of the corresponding item in Hedden & Zhang's experiment. My items are divided into 4 sets of 10 items each. Each set is a mix of the items in the same set from Hedden & Zhang, but from both their blocks. Items 21-24 are for familiarization with the new game mechanics, and no predictions are asked. Items marked with * are control items.

A.2 Instruction

A.2.1 In Dutch

Voordat het kind arriveert wordt het computerprogramma opgestart met als naam van de speler 'BLAUW'. Het laptopscherm wordt omlaag geklapt.

Goedemorgen, ik ben Liesbeth. Ben jij ...? (naam van kind)

We gaan straks een computerspelletje doen. Daarna vertel ik een paar verhaaltjes waar je vragen over moet beantwoorden.

Het laptopscherm wordt omhoog geklapt. In beeld is nu het parcours met auto, maar nog zonder knikkers.

Dit is het computerspel. Je gaat straks samen met de computer een autootje besturen. Jij bent blauw en de computer is geel. Als de auto op een blauwe kruising komt, mag jij bepalen welke kant de auto op gaat. Als de auto op een gele kruising uitkomt, mag de computer bepalen welke kant de auto op gaat. Samen met de computer bepaal je waar de auto uiteindelijk terechtkomt. Aan het eind van ieder spel krijg je een aantal knikkers. Klik maar op OK om het eerste spel te zien.

Proefpersoon klikt op OK. Nu komt het eerste item in beeld.

De auto rijdt alvast naar de eerste kruising. Je ziet nu op verschillende plekken knikkers liggen. De auto kan daar uitkomen, of daar, of daar. Je moet proberen de auto zo te besturen dat je zo veel mogelijk blauwe knikkers krijgt. De blauwe knikkers zijn voor jou en de gele knikkers zijn voor de computer. De computer probeert de auto zo te besturen dat hij zoveel mogelijk gele knikkers krijgt. De blauwe knikkers die je verzamelt, komen in de buis terecht links op het scherm. Als alle spellen geweest zijn, kijken we hoe vol de buis is. Als de buis vol is geraakt tot deze streep, dan mag je aan het eind, na de verhaaltjes, een sticker uitzoeken, maar alleen uit deze doos. Als de buis vol is tot deze streep, dan mag de sticker ook uit dit doosje komen, maar het hoeft niet. Je moet dus proberen zoveel mogelijk blauwe knikkers te verzamelen. Het maakt voor jouw score niet uit hoeveel gele knikkers te computer krijgt.

Heb je nog vragen? *(Beantwoord eventuele vragen)*

Dan mag je gaan spelen. Het autootje staat nu op de blauwe kruising. Dat betekent dat jij mag bepalen welke kant de auto op gaat. Dat staat ook onderaan: "Welke kant wil je op? Klik op een van de pijlen?" Kies maar.

Bij het vijfde item (eerste voorspelling):
Nu moet je eerst voorspellen wat de computer gaat doen. Welke pijl zou de computer kiezen als de auto op de gele kruising zou staan? Klik op die pijl. Daarna mag je weer gewoon zelf de auto besturen.

Als de training afgelopen is (er klinkt een geluidssignaal):
Nu wordt het moeilijker. Maar je kunt ook meer knikkers ineens verdienen. Klik maar op OK om te zien hoe het er nu uitziet. Er zijn nu drie kruisingen. De eerste is blauw, dus daar mag jij bepalen welke kant de auto op gaat. De tweede is geel, dus daar mag de computer bepalen waar de auto heengaat. En als de auto bij de laatste kruising komt, dan mag jij weer kiezen.

A.2.2 English translation

Before the child arrives the computer program is started with the player name 'BLUE'. The laptop is then closed.

Good morning, I am Liesbeth. Are you ...? (name of child)

We are going to play a computer game. After that I will tell you a few stories about which you have to answer some questions.

The laptop is opened. The program shows the empty track with car, but without marbles.

This is the game. Together with the computer you are going to drive the car. You are blue and the computer is yellow. If the care reaches a blue intersection, you may decide in which direction the car will go. If the car reaches a yellow intersection, the computer decides in which direction the car will goes. Together with the computer you determine where the car will end up. At the end of each game you will receive a number of marbles. Click on OK to see the first game.

The subject clicks on OK. The first item appears on the screen.

The car has already driven to the first intersection. You see there are marbles in different locations. The car can end there, or there, or there. You should try to drive the car in such a way that you get as many blue marbles as possible. The blue marbles are for you and the yellow marbles are for the computer. The computer tries to drive the care in such a way that he gets as many yellow marbles as possible. The blue marbles that you collect appear in the tube on the left of the screen. When all games have finished we will look how full the tube is. If the tube has been filled until this bar, then at the end, after the stories, you may pick a sticker, but only from this box. If the tube has been filled until this bar, you may also pick your sticker from this box, but you don't have to. So

you should try to get as many blue marbles as possible. It doesn't matter for your score how many yellow marbles the computer gets.

Do you have any questions? *(Answer questions if there are any)*

Now you may play. The car is at the blue intersection. That means that you may decide in which direction the car will go. That's what it says over there: "Where do you want to go? Click on one of the arrows?" Choose.

When the fifth item is reached (first prediction):
Now you have to predict what the computer will do. Which arrow would the computer choose if the car were standing at the yellow intersection? Click on that arrow. After that you may drive the car normally again.

When the training is finished (there will be a sound):
Now it gets more difficult. But you can also obtain more marbles. Click OK to see what the game looks like now. There are now three intersections. The first one is blue, so you get to decide in which direction the car will go. The second one is yellow, so the computer decides where the car will go. And if the car reaches the last intersection, you may decide again.

Appendix B

False belief task

B.1 Chocolate bar

See page 45 for the drawings accompanying this story and the English text.

B.1.1 In Dutch

Jan en Marie zijn broer en zus. Ze zijn in de woonkamer. Dan komt moeder thuis van het boodschappen doen. Moeder heeft chocola gekocht. Ze geeft de chocola aan Jan. Marie krijgt geen chocola, omdat ze stout is geweest. Jan eet wat van de chocola en doet de rest in de la. Hij geeft niets van de chocola aan Marie. Dat maakt Marie boos. Nu gaat Jan moeder helpen in de keuken. Hij gaat helpen afwassen. Marie is alleen in de woonkamer. Jan is in de keuken. Omdat Marie boos is op Jan, verstopt ze de chocola. Ze haalt de chocola uit de la en stopt het in de speelgoedkist.

Jan is bezig met afwassen. Hij gooit de fruitschillen weg in de vuilnisbak in de tuin. Door het raam ziet hij de woonkamer. Hij ziet dat Marie de chocola uit de la haalt, en in de speelgoedkist stopt. Marie ziet Jan *niet*.

Reality control question: Waar is de chocola nu?
1st order ignorance: Weet Jan dat Marie de chocola verstopt heeft in de speelgoedkist?
Linguistic control: Weet Marie dat Jan gezien heeft dat ze de chocola verstopte?

Jan is klaar met afwassen. Hij heeft honger. Nu wil Jan wat van zijn chocola eten. Jan gaat naar de woonkamer. Hij zegt: "Hmm, ik heb zin in chocola."

2nd order false belief: Waar denkt Marie dat Jan de chocola gaat zoeken?
Justification: Waarom denkt ze dat?

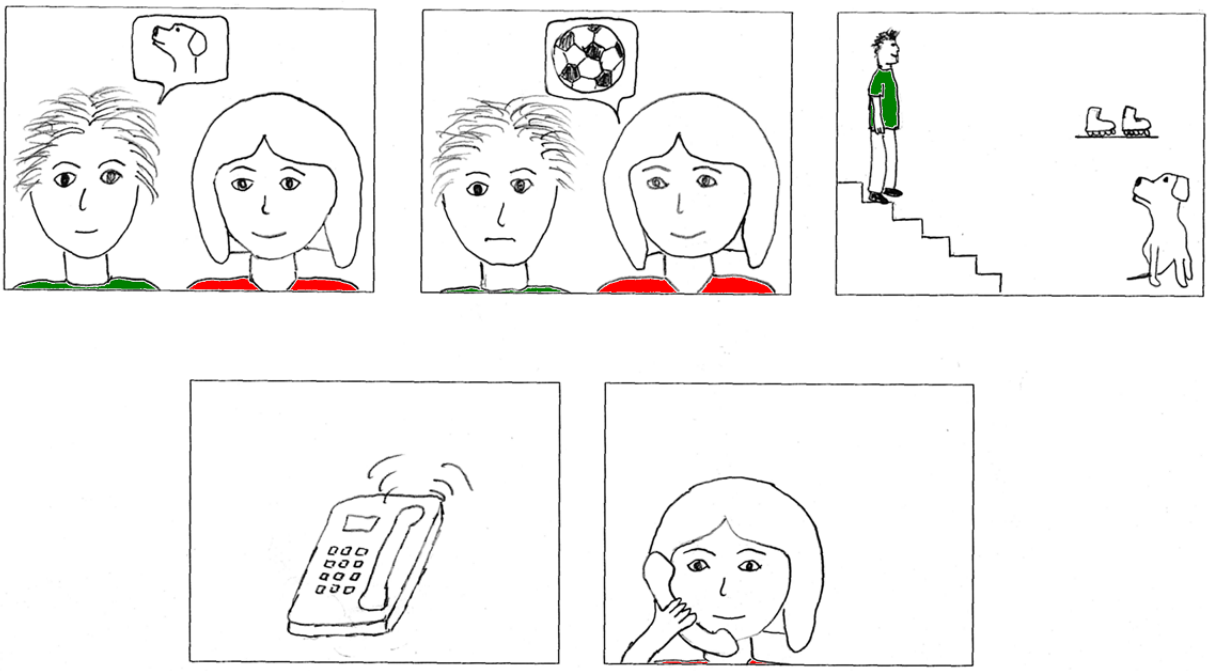


Figure B.1: The drawings accompanying the birthday puppy story.

B.2 Birthday puppy

A few changes were made to the story as written by Tager-Flusberg and Sullivan (1994). In the original story, mother tells Peter she got him a toy. I chose to use a more concrete object: a (soccer) ball. Also, the roller skates in the story were replaced with inline skates ('skeelers' in Dutch). The second-order ignorance question was omitted.

B.2.1 In English

Tonight it's Peter's birthday and Mum is surprising him with a puppy. She has hidden the puppy in the basement. Peter says, "Mum, I really hope you get me a puppy for my birthday." Remember, Mum wants to surprise Peter with a puppy. So, instead of telling Peter she got him a puppy, Mum says, "Sorry, Peter, I did not get you a puppy for your birthday. I got you a really nice soccer ball instead."

Reality control question: What did Mum *really* get Peter for his birthday?

Now, Peter says to Mum, "I'm going outside to play." On his way outside, Peter goes down to the basement to fetch his inline skates. In the basement, Peter finds his birthday

puppy! Peter says to himself, "Wow, Mum didn't get me a ball; she really got me a puppy for my birthday." Mum does *not* see Peter go down to the basement and find the birthday puppy.

1st order ignorance: Does Peter know that Mum got him a puppy for his birthday?
Linguistic control: Does Mum know that Peter saw the birthday puppy in the basement?

Now the telephone rings, ding-a-ling! Peter's grandmother calls to find out what time the birhtday party is. Mum tells grandma on the phone that she got Peter a puppy for his birthday, but that Peter doesn't know this. Then, Grandma asks Mum on the phone, "What does Peter think you got him for his birthday?"

2nd order false belief: What does Mum say to Grandma?
Justification: Why does Mum say that?

B.2.2 In Dutch

Vanavond is Peter jarig en moeder wil hem als verrassing een jong hondje geven. Ze heeft het hondje verstopt in de kelder. Peter zegt: "Mam, ik hoop dat je een hondje koopt voor mijn verjaardag." Denk eraan, moeder wil Peter verrassen. Daarom vertelt ze niet dat ze een hondje gekocht heeft, maar zegt ze: "Sorry, Peter, ik heb geen hondje gekocht voor je verjaardag. Ik heb mooie voetbal gekocht."

Reality control question: Wat heeft moeder *echt* gekocht voor Peter's verjaardag?

Nu zegt Peter tegen moeder: "Ik ga buiten spelen." Onderweg naar buiten gaat Peter naar de kelder om zijn skeelers te pakken. In de kelder ziet Peter het hondje voor zijn verjaardag. Peter zegt tegen zichzelf: "Wow, moeder heeft geen voetbal voor me gekocht; ze heeft een hondje gekocht voor mijn verjaardag." Moeder heeft *niet* gezien dat Peter naar de kelder is gegaan en het hondje gevonden heeft.

1st order ignorance: Weet Peter dat moeder een hondje heeft gekocht voor zijn verjaardag?
Linguistic control: Weet moeder dat Peter het hondje in de kelder gezien heeft?

De telefoon gaat, ring-ring. Peter's oma belt om te vragen hoe laat het verjaardagsfeestje is. Moeder vertelt aan oma over de telefoon dat ze een hondje voor Peter's verjaardag gekocht heeft, maar dat Peter dat niet weet. Dan vraagt oma aan moeder over de telefoon: "Wat denkt Peter dat je voor zijn verjaardag gekocht hebt?"

2nd order false belief: Wat zegt moeder tegen oma?
Justification: Waarom zegt moeder dat?

Appendix C

Excluded subjects

The table below[1] gives some information on the 19 subjects who were excluded from the second order analysis of the strategic game. The table lists the number of prediction errors in the last six training items. It also lists the number of rationality errors in the last six training items (actually, only in the items that were predicted correctly). The last column gives the proportion of incorrect actions at the last junction during the test phase. A proportion rather than a number was chosen because the number of times that this junction was reached is not the same for all subjects.

[1]This appendix refers to subjects by number. The first four adult subject were immediately excluded because of a subsequent change in instruction and rewards. The 27 remaining adult subjects have numbers 5 - 31, the 40 child subjects have numbers 33 - 72.

subject	**group**	**training phase**		**test phase, last junction**
		predictions errors	rationality erros	proportion of errors
20	adult	0	3	0.11111111
35	child	1	1	0.30769231
37	child	0	3	0.23529412
39	child	1	1	0.11764706
46	child	1	2	0.00000000
47	child	2	2	0.11764706
48	child	0	3	0.05263158
49	child	1	1	0.00000000
50	child	1	1	0.31578947
52	child	2	0	0.30000000
53	child	2	0	0.00000000
55	child	3	0	0.31578947
56	child	4	2	0.28571429
57	child	4	0	0.50000000
59	child	0	0	0.30769231
60	child	1	1	0.00000000
63	child	4	2	0.00000000
66	child	2	3	0.00000000
67	child	0	3	0.55555556
70	child	2	1	0.31578947