Reading experiments on higher-order social reasoning

Eva L. van Viegen

Juli 2014

Master's Thesis

Artificial Intelligence

Institute of Artificial Intelligence

University of Groningen, The Netherlands

First supervisor and reviewer:

Prof. Dr. Rineke Verbrugge (Institute of Artificial Intelligence, University of Groningen)

Other supervisors:

Dr. Ben Meijering (Institute of Artificial Intelligence, University of Groningen)

Dr. Jakub Szymanik (Institute for Logic, Language and Computation, University of Amsterdam)

Second reviewer:

Dr. Marieke van Vugt (Institute of Artificial Intelligence, University of Groningen)

Abstract

The ability to reason about other people's knowledge, belief, and intentions is called the theory of mind (ToM). To understand ToM better, two reading experiments were done. In the first experiment, participants were presented with stories about everyday situations. They were asked to memorize the story and afterwards answer higher order knowledge questions about them.

The first experiment indicated some interesting non-significant trends. Supposedly, this was due to a ceiling effect. Therefore, a second experiment was designed with a higher cognitive load. The cognitive load was increased by adding a higher/lower-game to the experimental setup. Now participants needed to memorize a number as well as the story while answering the questions.

The results of this second experiment indicated that the order of knowledge (OoK) influenced the total reaction times to the questions. However, a division of the total reaction times in a reading part and a decision part showed that OoK did neither influence the reading times or the decision times. The reading times were mainly influenced by the length of the questions. The decision times were influenced by the length of the question as well as by the self-reflexivity. The self-reflexivity is true whenever the question concerns the knowledge of another character about the participants' knowledge.

Table of contents

1 Introduction

The ability to reason about world facts, knowledge, and beliefs is well developed in humans. People reason every day, again and again, and there are even jobs based on solely this ability, not just a few, low-paid jobs, but many important ones, for example, jobs in politics, journalism, scientific research, and education. People performing these jobs transfer and combine information and especially in the case of scientific research, new world facts are found.

As reasoning about world facts and their implications on the knowledge and beliefs of people is so important, much research has been done in that area. Modeling other people's beliefs, knowledge and intentions is called having a theory of mind (ToM). It is related to folk, intuitive or commonsense psychology. These terms are slightly ambiguous (Stich & Ravenscroft, 1994), but for the purpose of this thesis they refer to an internal representation of human psychology. In that sense, using folk psychology is similar to having ToM.

Although ToM was presented as an absolute measure in the last paragraph, it is modeled to have different degrees. A zero-order sentence such as "there is an apple on the table" represents a fact about the world. A first-order attribution represents a person's belief about a fact, for example, "Alice *believes* that the apple is on the table". A second-order attribution represents a person's belief about another person's belief of a fact, as in "David does not believe that Alice believes that there is an apple on the table".

In this approach, different degrees of ToM are distinguished and experimental settings are designed where a certain degree of ToM is needed to pass the experimental task. Whenever a person, child, animal or computer passes the test, it is said to possess the corresponding degree of ToM. An example of a second-order belief task is when a participant sees two kids playing in a room. In the room are two boxes. Together the two kids put the ball into the left box. One kid leaves the room and the other kid secretly puts the ball into the other box. When the first kid comes back, the experimenter asks him where his playmate will start looking for the ball. The participant is asked to predict the answer to this question.

Another approach is to logically model reasoning about world facts. The simplest models, as in propositional logic, only include logical inferences about world facts. More complicated models, as in epistemic logic, allow reasoning about the knowledge of others. Even more complex models, as in dynamic epistemic logic, allow world facts and knowledge of people to change.

Researching what degree of ToM tasks humans, children at different ages, different kinds of animals, and/or computers pass does not offer us more knowledge about how

ToM reasoning is actually executed. How do humans reason about others? Nevertheless, such ToM tests might give us some indication about what kinds of higher-order reasoning tasks are most difficult.

Human adults are better in reasoning about other people's thoughts and intentions than human children and animals (Apperly, 2011). So good, that they are believed to have a ToM. The distinction between reacting to behavior and choosing behavior according to ToM reasoning might not be that strict (Gallese, 2007) in both animals and humans.

The imperfections in the reasoning of human adults were explored in order to get a better understanding of the processes needed to pass ToM tasks. In other words, what variables influence the accuracy and speed when human adults pass or fail ToM tasks?

To track the imperfections in adult reasoning, participants had to answer questions related to stories. Both the stories and the questions differed in difficulty. The difficulty of the stories was manipulated with their underlying Kripke structure; the difficulty of the questions with their length, order of knowledge, and self-reflexivity.

The difficulty of the questions was measured by determining the accuracy of the answers to the questions and the total reaction times needed to answer the questions. In addition, the reading times and the decision times to the questions were determined to distinguish between different stages of the process of answering: reading and processing the question meaning on one hand, and formulating and determining the answer on the other hand. The difficulty of the story was derived from the measurements on the questions related to the stories.

In Chapter 2, an overview is given of the relevant literature on ToM, the difference between simulation theory and theory-theory, epistemic logic, how different agents possess different knowledge, and self-reflexivity. In Chapter 3, this theory is applied to the creation of stimuli for the experiment. In Chapter 4, the theoretical background is discussed regarding the methodologies used in the experiment. In Chapter 5, the research question and hypotheses are discussed. Chapters 6 and 7 discuss experiment 1 and 2 respectively. In Chapter 8, a general discussion is provided. In Chapter 9, recommendations are made for further studies. In Chapter 10, some concluding remarks are given along with a short summary of this thesis. In Chapter 11, a list of abbreviations is provided.

2 Theoretical framework

To investigate everyday reasoning of human adults, I performed some reading experiments. Participants needed to read stories and answer first-order and second-order reasoning questions about them. For the interpretation of the results, it is important to understand the theoretical context around the experiment. Therefore, some background is provided for the origins of theory of mind and the corresponding orders of reasoning. Also, two main theories about the way people reason about others – the simulation theory and the theory-theory – are explained and compared shortly.

As the resulting knowledge after reading the stories can be described logically with Kripke models, I will shortly rehearse some epistemic logic. Furthermore, I will explain with some examples how knowledge can differ between agents within stories. Finally, I will explain the concept of self-reflexivity. This is a new term reflecting whether the actual knowledge of a participant conflicts with the knowledge of another character about the knowledge of the participant himself.

2.1 Theory of mind

In 1978, (Premack & Woodruff, 1978) formulated a definition of theory of mind (ToM): "An individual has a theory of mind if he imputes mental states to himself and others". These mental states may be desires, beliefs, intentions or knowledge. This definition divides individuals in two groups, ones with a theory of mind and ones without it. This way of thinking has often been shown in research questions, for example: "Does the chimpanzee have a theory of mind?" (Premack & Woodruff, 1978), "Does the autistic child have a theory of mind?" (Baron-Cohen, Leslie, & Frith, 1985).

However, ToM does not seem to be an absolute ability. There are several aspects to this ability that children develop at different ages. One aspect is that of pretending. To assign beliefs and desires to others, it is needed to imagine other facts. Children show this from the age of two, when they are able to pretend that, for example, a banana is a telephone, even though they do know it actually is a banana (Leslie, 1987). Another aspect of the ability to apply ToM that develops around the age of 3-4 is the ability to understand that other people may have beliefs other than one's own beliefs and that those beliefs may be false.

The most famous task to prove that someone has ToM is the false belief task. A false belief task can be described by the following paradigm. A subject and some other person observe some state x. In the absence of this person, the state suddenly changes from x to y. The subject does now believe that y is the case, while the other person still thinks x is the case (Wimmer & Perner, 1983).

However, people, animals, children, and other agents cannot be divided into simply two groups: one with and one without a ToM. One reason for this is that children pass some false-belief tasks at 13 to 15 months (Onishi & Baillargeon, 2005), but still fail critical belief reasoning tasks before 3 or 4 years of age (Wimmer & Perner, 1983; Apperly & Butterfill, 2009). In other words, different aspects, like knowledge attribution and false-belief reasoning, of ToM in children manifest themselves at different ages (Leslie, 1987).

Also, research on non-human animals does not prove conclusively whether there are animals having a ToM (Penn & Povinelli, 2007; Heyes, 1998). This inability suggests that there are still many uncertainties about ToM. Part of this inability is due to the fact that indirect measures are used to determine whether animals (and small infants) have ToM (Apperly, 2011, chapter 3). This is because, we cannot simply ask them about the knowledge of other people and animals, as their language skills are not sufficient. Therefore behavior that was thought to suggest that a chimpanzee may have a ToM (Premack & Woodruff, 1978), may be explained by associative learning or other non-mental processes (Heyes, 1998).

There are several assumptions used in the definition of the different degrees of ToM. One of them is: all individuals have positive introspection on their knowledge. So “I know p” is equivalent to “I know that I know p”. However, “I know p” is not equivalent to “Peter knows that I know p”. However, “I know that Peter knows p” is equivalent to “I know that Peter knows that Peter knows p”, as the positive introspection is assumed for all agents.

For example: When David thinks that the chocolate bar lies on the table, he does not use ToM. However, when David thinks that Nina thinks that the chocolate bar lies on the table, David uses first-order ToM, because he thinks something about someone else’s knowledge. This can be repeated recursively, to get higher orders of ToM. For example: David thinks that Nina thinks that Peter thinks that Erik thinks that the chocolate bar is on the table. In this example, David expresses third-order ToM, and we make a fourth-order attribution to David.

Other examples: “David thinks that he thinks that Peter thinks that the chocolate bar lies on the table”. David expresses here first-order ToM and not second-order ToM. Same goes in the following example: “David thinks that Peter thinks that Peter thinks that the chocolate bar lies on the table”. As Peter knows what Peter knows, this would be logically the same as: David thinks that Peter thinks that the chocolate bar lies on the table. And therefore, David expresses just first-order ToM and not second-order ToM. However, in the following example this does not apply: “David thinks that Peter thinks that David thinks that the chocolate bar lies on the table”. In this example, David expresses second-order ToM, as David does not think about his own knowledge anymore, but about what someone else thinks about David’s knowledge.

At first sight, it is tempting to draw conclusions about the degree of ToM which children at a certain age, animals, and adults can effectively use. However, for adults it has been shown that they can solve some second-order belief tasks, but have more difficulty with others (Birch & Bloom, 2007).

Another conclusion that has often been drawn is that having a ToM makes us different from other animals. Animals are thought to read each other's behavior and adjust their behavior. For example, deer start running as they see their flock mates running. The distinction between reacting to behavior and choosing behavior according to ToM reasoning might not be that strict (Gallese, 2007).

2.2 Theory-theory vs simulation theory

There are different ideas about how people interpret other people's behavior. They can be divided into two different streams: simulation theory and theory-theory.

Theory-theory is based on a commonsense ToM and it is linked to folk-psychology (Gallese & Goldman, 1998). The idea is that people have some explanatory laws that link behavior to thoughts. These laws, or in other words theories, are constantly updated (Gopnik & Wellman, 2012). For these updates, both information from our own explorations in the world and information obtained from watching others may be used (Gopnik & Wellman, 2012).

These seem really simple, straightforward laws, but there is a problem with this approach. How many laws are needed and how are they represented? If all human behavior needs to be explained and people seem quite able to do so, an unbounded number of rules are necessary.

The idea behind the simulation theory is that the mental states of others are matched with a person's own behavior in a certain situation (Gallese & Goldman, 1998). As the behavior of the other person is mirrored, the representation of that person's behavior is assumed to be the same as your own.

Research in monkeys has shown that mirror neurons exist (Gallese & Goldman, 1998). Mirror neurons respond to a certain action regardless whether the action is perceived or executed. In fMRI studies on humans, groups of neurons have been found that respond both when a subject performed an action and when a subject perceived the same action (Gazzola & Keysers, 2009). As this provides some evidence of the representation of simulation theory in the human brain, it is likely that at least some human reasoning processes work with simulation.

In the next section, Kripke models and epistemic logic are discussed.

2.3 Kripke models and epistemic logic

Most readers may be familiar with epistemic logic and Kripke models. However, the main points, necessary to understand the analysis of the stories and questions of the experiments, are repeated in this section. See (Meyer & van der Hoek, 2004; van der Hoek & Verbrugge, 2002) for more extensive introductions to epistemic logic.

Epistemic logic may be viewed as an extension of modal logic in order to represent knowledge of several agents instead of plain world facts. The relationship of this logic with Kripke models will be discussed in this section, along with the S5 axiom system.

Epistemic logic may be viewed as an extension of modal logic in order to represent the knowledge of agents. Therefore, the language of epistemic logic needs to include a way to represent this knowledge information in addition to the representation of facts. This results in the following definition of the language for epistemic logic. Note that part (iii) introduces the new idea of knowledge.

Given that P is a set of atoms and A a set of agents, conveniently numbered 1 to m. The set $\mathcal{L}_{\mathbf{K}}^{\mathrm{m}}(\mathbf{P})$ of epistemic formulas $\varphi, \psi, \ldots$ over A is the smallest set closed under:

(i)If $p \in \mathbf{P}$ then $p \in \mathcal{L}_{\mathbf{K}}^{\mathrm{m}}(\mathbf{P})$.

(ii) If $\varphi, \psi \in \mathcal{L}_{\mathbf{K}}^{\mathrm{m}}(\mathbf{P})$ then $(\varphi \wedge \psi), \neg\varphi \in \mathcal{L}_{\mathbf{K}}^{\mathrm{m}}(\mathbf{P})$.

(iii)If $\varphi \in \mathcal{L}_{\mathbf{K}}^{\mathrm{m}}(\mathbf{P})$ then $\mathrm{K}_i\varphi \in \mathcal{L}_{\mathbf{K}}^{\mathrm{m}}(\mathbf{P})$, for all $i \in A$.

Definition 1: Epistemic formulas (Meyer & van der Hoek, 2004)**.**

However, this language does not provide any truth assignments or reasoning mechanisms. It just describes all possible sentences in the logic. To reason with this logic, it is important to notice that the truth values of sentences may differ among states. Even among states with the same truth assignments to their atoms. (To reason about knowledge, it is important to take into account the truth values of sentences at other states) Some facts are always true (like tautologies), but many others depend on the 'world'. Kripke models are a way to represent possible states. The formal definition is as follows:

A Kripke model $\mathbb{M}$ is a tuple $\langle \mathrm{S}, \pi, \mathrm{R}_1, \ldots, \mathrm{R}_{\mathrm{m}} \rangle$ where:

(i) S is a non-empty set of states,
(ii) $\pi: S \rightarrow (P \rightarrow [t, f])$ is a truth assignment to the atoms per state,
(iii) $R_i \subseteq S \; x \; S \; (i = 1, \ldots, m)$ are the possibility relations.

Definition 2: Kripke models (Meyer & van der Hoek, 2004)**.**

In other words, a Kripke model consists of several states, each with a specific truth assignment in each state, for the propositional atoms. Between every combination of two states, there may be a accessibility relation for one or more agents. When there is an accessibility relation from one state to another state for an agent, this means that for the agent, the other state is consistent with his information in the original state. A Kripke world is a Kripke model with one state assigned as the real one.

Possibility relations may originate and end in the same state. Some logical frameworks demand certain restrictions on possibility relations, for example that they are reflexive, transitive, and symmetric. One of the famous axiom systems is S5. It consists of five axioms (Definition 3) and two derivation rules (Definition 4) that will be discussed one by one.

(A1) All (instances of) propositional tautologies
(A2) $\big(K_i\varphi \wedge K_i(\varphi \rightarrow \psi)\big) \rightarrow K_i\psi$ for $i = 1, \dots, m$
(A3) $K_i\varphi \rightarrow \varphi$ for $i = 1, \dots, m$
(A4) $K_i\varphi \rightarrow K_iK_i\varphi$ for $i = 1, \dots, m$
(A5) $\neg K_i\varphi \rightarrow K_i\neg K_i\varphi$ for $i = 1, \dots, m$

Definition 3: axiomas of S5 (Meyer & van der Hoek, 2004)**.**

To represent the knowledge of an agent, it gets basic facts to work with and axioms that are always true. The first axiom is that instances of all propositional tautologies are true. An example of a propositional tautology is $(p \vee \neg p)$. As p can either be true or false, either p or $\neg p$ is true, and therefor this formula is always true.

The second axiom is called Modus Ponens; it states that when an agent knows φ and the same agent also knows $\varphi \rightarrow \psi$, then the agents also knows ψ. So this formalizes that an agent is able to make inferences. This represents the idea of logically capable agents. They are not some kind of database, but they are actually capable of reasoning with the facts they know.

The third axiom is that known facts are true. This axiom means that agents are sane, so when they know something, it should be true in the actual world. However, in the real world this does not always seems to be the case. For example, before Galileo, sane people on earth knew that the sun was circling the earth. In the present days, most people know that the earth is actually circling the sun. Is it fair to say that those people before Galileo just believed that the sun was circling the earth? Do we really know that the earth is circling the sun? How much evidence is necessary before beliefs turn into knowledge?

The fourth is called positive introspection, which means that an agent knows that she knows something. This is true for all agents, so when an agent is reasoning about

another agent, it also assumes it has positive introspection. For basic facts this assumption seems straightforward, however, it also applies to inferred knowledge. For example, an agent may not know the answer to a difficult equation at first, but will know it after calculating and reasoning a while.

The fifth – negative introspection – is left out in some frameworks; it means that an agent knows that she does not know something. Again this would also apply when reasoning about other agents. But more importantly, in this case, the calculation costs to determine you do not know something are even higher than in the positive introspection case. This is probably one of the main reasons it is sometimes left out as an axiom.

$$\text{(R1)} \quad \frac{\varphi \quad \varphi \rightarrow \psi}{\psi}$$

$$\text{(R2)} \quad \frac{\varphi}{K_i \varphi} \text{ for } i = 1, \dots, m$$

Definition 4: derivation rules for S5 (Meyer & van der Hoek, 2004)**.**

The derivation rules of S5 (Definition 4) are called modus ponens (R1) and necessitation (R2). The combination of axioms and derivation rules allow reasoning. A formula φ is provable when it is an instance of an axiom of can be derived from an axiom using any number of derivations. Where a derivation is an application of one of the derivation rules on an axiom or another formula that was already derived from an axiom with one or several derivation rules.

Besides knowledge, there are a few other important concepts in epistemic reasoning: belief, common knowledge and common belief. For this thesis, the distinction between knowledge and belief is not relevant. When I look at human reasoning, it is often not clear whether people know or whether they think something. A common definition is that someone knows a fact when the fact is true, the person beliefs the fact and has reason to do so. This definition is not sufficient to prove knowledge (Gettier, 1963). However, for the purposes of this thesis, it is clear enough to grasp some of the conceptual differences between knowledge and beliefs.

Common knowledge is easy to explain: something is common knowledge when everyone knows this and everyone knows that everyone knows this and this is continued recursively. Therefore, logically, common knowledge is often hard to prove and therefore often impractical to use.

Examples of Kripke models applied to the stories used in the experiments of this thesis may be found in Section 3.1. With these examples, the use of Kripke models becomes even clearer, especially with the theoretical background in this section.

2.4 Fooling other agents

The notion of pretending depends on the ability to have two theories about one's own beliefs, the one corresponding to the world facts and the one corresponding to the pretended world (Leslie, 1987). In the following sections, three different forms of pretending are discussed, along with the theory of mind representations needed in both the manipulator and the interpreter.

The most important form of pretending for this thesis is changing world facts. To illustrate this sort of pretending let us consider the following example. Levi and Nina go out to eat ice cream. They order vanilla ice cream. But when Levi leaves for the bathroom, Nina changes the order to banana ice cream. Now, Nina knows that they will get banana ice cream and more importantly, she knows that Levi thinks they will get vanilla ice cream. Levi is not aware that they will get banana ice cream and has no reason to consider other options than vanilla.

Even within this simple form of pretending, the manipulator, Nina, needs to be able to have a theory about Levi's knowledge for changing world facts really to constitute pretense. When world facts were not changed in order to trick someone, changing world facts does not constitute pretense.

Another form of pretending is telling someone that the world facts are different from what they actually are. An example is that Levi tells Nina that he bought a book for her birthday, while he actually bought her a video game. Now, the assumption is that Nina the interpreter adds to her interpretation that Levi probably tells the truth.

This last form of pretending may also be called lying. Lying is something else than simply not telling the truth. There are two additional conditions before something is a lie. The speaker should believe that the utterance is false. Furthermore, she must intend the utterance to be taken by the listener as truth (van Ditmarsch, van Eijck, Sietsma, & Wang, 2010).

Often, announcements, regardless whether they are private or public, can only be made when the announcer believes the announcement himself. In some frameworks it is even not possible to make announcements unless the announcer knows something. So these frameworks exclude lies as defined above.

The last form of pretending is not telling someone that a certain world fact is different than the person probably believes. This seems similar to the first form of pretending, but there is a subtle difference. In the first form, the world fact is changed by the manipulator, while in this form, the world fact is changed by a third party and the interpreter does not know about this change. However, this world fact does not have to be explicitly changed, the fact may also be unknown to the fooled agent.

An example of this kind of pretending is present in the following story. Nina and Levi are in a bar together. They decide to drink some vodka. When Levi leaves for the bathroom, the bartender replaces the vodka with water, because she thinks Levi is too drunk. When Levi returns, you and the bartender pretend there is still vodka in Levi's glass. Here, you are a manipulator of the third form, whereas the bartender is a manipulator of the first form.

2.5 Self-reflexivity

Young children are known to have difficulties handling situations where they have to imagine the knowledge of others (Leslie, 1987). They especially find it difficult to reason about beliefs that conflict with reality (false beliefs) (Wimmer & Perner, 1983). With precise measures, adults were shown to have similar difficulties when reasoning about false beliefs (Birch & Bloom, 2007; Lin, Keysar, & Epley, 2010).

Another experiment by (Samson, Apperly, Braithwaite, Andrews, & Scott, 2010) showed participants images of a room with an avatar in it. There could be dots on both the wall the avatar was facing and the opposite wall. The participants could see all dots, but the avatar could only see the dots in front of him. So when there were dots on the wall behind the avatar, the number of dots the avatar could see and the number of dots the participant saw differed. If all the dots were on the wall the avatar was facing, the avatar would see the same number of dots as the participant. It was shown that when the number of dots the participant could see differed from the number of dots the avatar could see that the participants made more mistakes and needed more time to think.

This experiment has been replicated with people in different age groups, including adults (Surtees & Apperly, 2012). Within this experiment, no differences were found in the egocentrism effects. Participants of all age groups showed more errors and longer reaction times when asked to evaluate the avatar's perspective than when asked about their own perspective. So the improved ability of adults compared to children on theory of mind tasks might not have a structural grounding.

When people reason about what other people belief about their knowledge and beliefs, they know what their own knowledge and beliefs are. In other words, they undoubtedly know the reality. On the other hand, when people reason about what other people belief about another person's beliefs, they do not always know the reality. For this reason, it is relevant to distinguish between those two cases.

In this thesis, the term self-reflexivity is used for situations where people need to reason about what other people think of their own knowledge. The question "Does the government know that you paid your taxes?" requires self-reflexive reasoning behavior. Whereas the question "Does the government know that Obama paid his taxes?" does not require self-reflexive reasoning. In the first question, the responder knows whether he

paid his taxes, but in the second question, the responder does not know whether Obama paid his taxes. The definition of self-reflexivity does not have to do with the ability of people to reflect upon their own behavior, which is another definition of self-reflexivity that does not apply for this thesis.

2.6 Summary

Some researchers say that humans distinguish themselves from other animals because they have a theory of mind. This is the ability to recognize the mental states of others and reason with this knowledge to predict other people's behavior. There is much controversy about the question how this ability is represented in the human brain.

Theories of how the "theory of mind"-ability is represented within the brain can be divided into two different main streams: simulation theory and theory-theory. Behind simulation theory lies the idea that people imagine themselves in other people's shoes and determine what they would do in that situation. Theory-theory is framed in terms of rules that can be applied to a certain situation.

In my thesis, I will use theory-theory to analyze my experimental results, but this does not mean I strongly favor this above simulation theory. The representations in terms of rules made it easier to test my hypothesis and allowed the use of the well-known Kripke models for the representation of knowledge.

To display the use of the human theory of mind behavior, some notions of pretending were discussed. These notions may be caused by changes in the real world or by telling lies and refraining to tell relevant truths. However, in all these forms an accurate prediction of the other person's knowledge is important.

The term self-reflexivity was introduced to distinguish between situations where people reason about other people's beliefs about facts or other people's knowledge and situations where people reason about other people's belief about the beliefs of the person himself. In the first case, the person does not necessarily know the reality, whereas in the second case he does.

3 Applying theory to the stimuli for the experiment

In this chapter, the main stimuli for the experiments are discussed: the stories and the questions. Participants needed to read and memorize the stories and answer the corresponding questions. In Chapter 2, different degrees of theory of mind, self-reflexivity and the way knowledge may differ among agents have been discussed. It is important to understand how these theories are integrated in the stories and questions.

This chapter starts with a section about the stories, as the content of the stories is necessary to understand and correctly evaluate the questions. After this, the questions are discussed, with an explanation on how and why they are suited to test some aspects of the theoretical framework discussed in Chapter 2.

3.1 Stories

One of the biggest challenges of this thesis was to construct the stories for the experiments. Apart from the fact that these stories should result in a well-defined belief model for the characters, they also needed to be fluent. Participants should not have any suspicion on what the goal of the experiment was. My first inspiration for the stories in this thesis is the famous chocolate bar story (Hogrefe, Wimmer, & Perner, 1986).

All stories semantically differed, but were structurally divided into three groups. These three groups all had a different Kripke model representing the beliefs of the characters at the end of the story. The Kripke models served two main purposes. First they allowed ambiguity-checking of the stories. In this way the answers to the stories could be determined conclusively. Second, it could be tested whether the complexity of the knowledge model influenced the speed of answering the questions.

During the construction of the stories, care was taken to make the stories both unambiguous and fluent. They needed to be unambiguous, in order to make distinctions between correct and incorrect answers; and fluent, because I was afraid that participants may otherwise evaluate the questions to the story in a mathematical and logical way, rather than in the way they do in everyday life. This combination turned out to be difficult.

The well-defined belief models were represented with Kripke models and served two main purposes. First they allowed ambiguity-checking of the stories. In this way the answers to the stories could be determined conclusively. Second, it could be tested whether the complexity of the knowledge model influenced the speed of answering the questions.

3.1.1 Story structure 1

This story structure has the simplest Kripke model with just two states. However, for this Kripke model, originally two different story structures existed. As those seemed to have no statistical differences, I decided to combine them when interpreting the experimental results. However, the original distinction is still explained in this section. Here follows the first example story.

Imagine that you (character A) are with Levi (character C) in his living room. Nina (character B) enters with a chocolate bar for Levi, because he had his birthday. Levi puts the chocolate bar on the table. Levi leaves the room to do groceries, Nina stays with you. You decide to hide the chocolate bar behind the closet. Nina sees that you hide the chocolate and you see her surprised look. Then you are tired and go home.

Story 1a

Within this story there are three characters. One of the characters refers to the participant ("you"); the other two characters were Levi and Nina. These same three characters were used throughout the experiment. The participant himself is part of the story to prevent ambiguity in the questions. This is explained further in Section 3.2.

The questions about this story inquire about the participants' knowledge about the whereabouts of the chocolate. This results in two truth-values for Kripke states: one with the chocolate on the table and one with the chocolate behind the closet. For this story, these two truth values result in two states, named 'Table' and 'Closet' in Figure 1.

Levi thinks that the chocolate is on the table and he thinks that the other characters think this too, as there was no reason given in the story to think otherwise. This is sometimes called the inertia principle (Stenning & van Lambalgen, 2007). So he only considers the state where the chocolate is on the table. The arrows in Figure 1 represent the belief-accessibility relations of the characters in the story.

Both the participant (A) and Nina (B) are aware of the knowledge and belief of Levi (C). However, they know for themselves that the chocolate is behind the closet. This results in the Kripke model in Figure 1. The reflexive arrows from state s1 to itself for the participant (character A) and Nina (character B) may seem strange, but remember that these represent the possible belief of Levi (character C) of the beliefs of the participant and Nina.

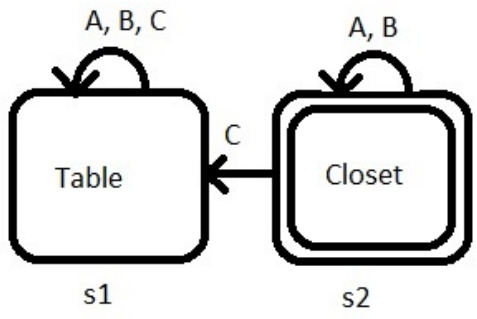


Figure 1: Kripke model structure 1

However, there are probably more factors influencing the comprehension difficulty of the story. In Story 1a there is only one change in beliefs in the story. At first, all characters believe that the chocolate is on the table. Then the chocolate is moved. Then, two of the three characters believe that the chocolate is behind the closet.

Another kind of story was constructed, based on the same Kripke model as shown in Figure 1. However, for this story, state s1 represents eating plain vanilla cake and state s2 represents eating chocolate cake. In this story, there are two changes in beliefs. An example story of this kind is offered here.

Imagine you (character A) are out with Levi (character C) and Nina (character B) for some cake. As you are all broke, you decide to eat plain vanilla cake. You did not tell Nina and Levi you just got your wage. You decide to treat them with some nice chocolate cake. When you arrive at the table with the chocolate cake, Levi has just gone away to get study books. Nina looks really happy and greedy at the chocolate cake and takes a bite right away.

Story 1b

In Story 1b, the questions inquire about the knowledge and beliefs of the characters of the food they are going to eat. In this story, Levi thinks that they are going to eat plain vanilla cake. The participant and Nina think that they are going to eat chocolate cake. So, this story results in the Kripke model in Figure 1.

This story is structurally a little different from the first one as there are two separate state changes. At first, all three characters think they are going to eat plain vanilla cake. Then the participant (character A) changes the plan. After this, only the participant thinks differently. Then the participant tells Nina (character B) about the chocolate cake. And only after this belief change, the situation is similar to that in Story 1a.

However, there were no statistically significant differences in difficulty found between stories like Story 1a and stories like Story 1b. Therefore, both these stories will be considered to be of story structure 1 in the presentation of the experimental results.

3.1.2 Story structure 2

Story structure 2 is a little bit more difficult than story structure 1, as the corresponding Kripke model contains three different states. Also, within this structure, some characters are actively pretending (lying) about the state changes, whereas in story structure 1, the uninformed characters were just not present. The example story is about whether a glass contains water or tequila.

Imagine you (character A) sit with Nina (character B) and Levi (character C) in a bar. You decide to drink some tequila. Levi goes to the toilet and you and Nina replace the tequila with water. Levi obviously cannot see that it is water, but you whisper this information into his ear. Nina has no idea you told Levi about the change.

Story 2

In Story 2, the participant knows that there is water in the glass and that both Nina and Levi know this too. However, Nina does not know that Levi knows that there is water in the glass. And the participant knows that she does not know this. This means that the accessibility relations are different for the participant than for Nina and therefore it cannot be represented within one state. Therefore the Kripke model has two different states in which the glass contains water.

However, also a state where the glass contains tequila is necessary, although no character actually believes that the glass contains tequila. This is because Nina thinks that this state is a possible state for Levi. This results in the Kripke model in Figure 2.

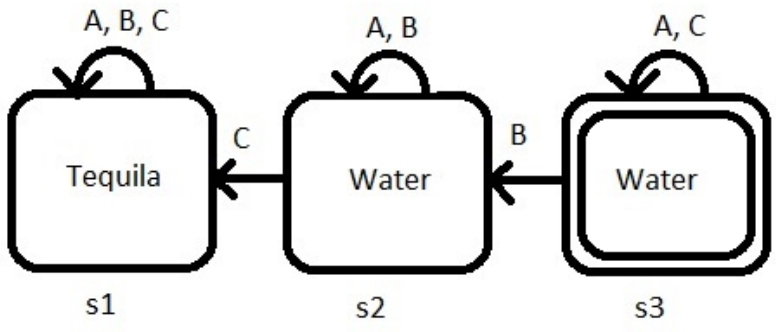


Figure 2: Kripke model structure 2

3.1.3 Story structure 3

Story structure 3 is similar to story structure 2. The biggest difference is that in this structure, the participant is temporarily unaware of the actual state. The questions to the example story, Story 3, are about the time of the tiger feeding. The corresponding Kripke model, displayed in Figure 3, is similar to that of story structure 2, but the characters A, B, and C are switched.

Imagine you (character A) go with Nina (character B) and Levi (character C) to the zoo[1]. You have an appointment to feed the tigers at five. You go to the insect house on your own, without Nina and Levi. After that you incidentally meet Nina. She tells you Levi changed the feeding time to three o'clock. You two keep walking together. Levi is nowhere in sight.

Story 3

In Story 3, all characters know that the tigers were supposed to be fed at five. However, Levi thinks that the participant does not belief this. Therefore, the accessibility relations differ between Levi and Nina.

The Kripke model in Figure 3 seems structurally equal to structure 2 (Figure 2), but there is a difference in the accessibility relations. This originates from the fact that the participant is represented by 'A' and Nina and Levi were represented by either 'B' or 'C'. This difference is important for the purposes of this thesis as in structure 3 Nina or Levi has an incorrect belief about the beliefs of the participants; whereas in structure 2 Nina or Levi has an incorrect belief about the other (not the participant). This fact changes the answers to the questions belonging to the story, classifying it as different structures.

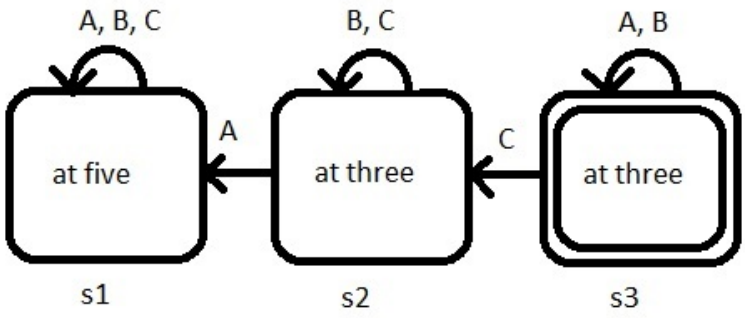


Figure 3: Kripke model structure 3

3.2 Questions

Every story contained three characters: Nina, Levi, and the participant ("you"). The questions were about the beliefs of these three characters and had the same structure: (Does X think that) (Y thinks that) Z thinks that …? For Story 1a (on page 17, repeated below) about chocolate, the dots were substituted by either "the chocolate is on the table" or "the chocolate is behind the closet". These question endings corresponded to the atoms used in the Kripke models of the stories.

Imagine that you (character A) are with Levi (character C) in his living room. Nina (character B) enters with a chocolate bar for Levi, because he had his birthday. Levi puts the chocolate bar on the table. Levi leaves the room to do groceries, Nina stays

[1] This sentence seems weird in English, but in the original Dutch version it is grammatically correct.

with you. You decide to hide the chocolate bar behind the closet. Nina sees that you hide the chocolate and you see her surprised look. Then you were tired and went home.

The characters in the questions are used to define the question structures. Therefore, the characters are represented by the letter A, B, and C, whereas the letters B and C are interchangeable. The letter A represents the participant; "you" in the question, the letters B and C may either represent "Nina" or "Levi", but not the same character. The first question in this section could therefore be represented by BA or CA, but as the first alphabetically comes first, that one was used. Every question contained one, two, or three characters.

The six question structures used were: B, AB, BA, BC, ABA, and ABC. Those structures differ in Order of Grammar, Order of Knowledge, and/or Self-Reflexivity. The question structures in combination with the story structures differed in the number of states. All these factors were discussed separately.

3.2.1 Order of Grammar

The questions are of the form: "Does X think (that Y thinks) (that Z thinks) that the chocolate bar is on the table?" Obviously, this question becomes longer when a character is added. The Order of Grammar (OoG) was used a measure of the length of the questions. It is calculated by simply counting the number of occurrences of characters in the question, which is equivalent to the number of occurrences of "that".

The question "Does Nina think that you think that the chocolate bar is on the table?" has OoG two. The question "Do you know that the chocolate is behind the closet?" has OoG one. The factual question "Is the chocolate behind the closet", would have OoG zero, but factual questions were not used in this experiment. The question with structure BCB would have OoG three: "Does Nina think that Levi thinks that Nina thinks that …?"

3.2.2 Order of Knowledge

In this thesis, the Order of Knowledge (OoK) is used, to name the degree of Theory of Mind (ToM). Because participants need to answer questions to a story, the OoK of a question is defined as the degree of ToM necessary for the participant to parse and answer that question. When the OoK of a question is zero, the question is about the participant's factual knowledge: "Does the chocolate bar lie on the table?" or "Does X think that the chocolate bar lies on the table?" Both these questions have OoK zero as there is no need for the participant to infer knowledge of others. As positive introspection and veridicality is assumed for the participant and the other story characters, there is no inferential difference between the two questions.

Another example question is: "Does Nina think that the chocolate bar lies on the table?" This question has OoK one, as the participant needs to have a model about Nina's knowledge. The OoK may increase recursively by adding "Do you think that" or "Does

Nina/Levi/Peter/etc think that" at the beginning of the question. However, due to the assumed positive introspection, adding those phrases does not always increase OoK.

For example, the question: "Does Levi think that Nina thinks that the chocolate bar lies on the table?" has OoK two. In contrast, both the questions: "Do you think that Nina thinks that the chocolate bar lies on the table?" and "Does Nina think that Nina thinks that the chocolate bar lies on the table?" have OoK one. In the first case, because the participant is introspective; and in the second case because Nina is introspective and veridical about own beliefs.

From the examples in the last section one might infer that only the number of different characters is important to determine the OoK. However, the total order of characters matters too. For example, the question: "Does Nina think that you think that Nina thinks that the chocolate bar is on the table", contains only two distinct characters but has OoK three.

3.2.3 Self-Reflexivity

Self-reflexivity (SR) in this thesis is used for situations where a character needs to make inferences about what someone else believes about his/her own beliefs. In contrast to inferences made about the beliefs of someone else about facts or other characters' beliefs. In the first case, the character knows the reality, because he is aware of his own beliefs. In the second case, the character does not always know what the reality is.

In questions that ask solely about facts, or about other characters' knowledge about those facts, the SR is always "No". The same is true when the participant is directly asked about those facts and other characters' knowledge. So the question: "Do you think that Nina thinks that *Levi* thinks that the chocolate is on the table?" has SR "No".In the question: "Does Nina think that *you* think that the chocolate bar is on the table?" the SR is "Yes". This is because in this case you are reasoning about your own belief and what Nina thinks your belief is.

Self-reflexivity could also be ranked 0, 1, 2, etc. but in that case the questions would become really large. For example, this question would have self-reflexivity of order 2: "Does Nina think that *you* think that Levi thinks that *you* think that the chocolate bar lies on the table?" However, these kinds of questions were not used in the experiments of this thesis. The main reason for this was that they are so long that the participants would be pushed to answer these questions in a formal/logical way. And also, the number of questions per story would have become so large that memory issues might have influenced the results of the experiments.

3.2.4 Number of states

The number of states that a question 'visits', is not just linked to the question structure, but also to the associated story. In Section 3.1, it was explained how the different story

structures led to different Kripke models. In this section, one possible way of reasoning of the participant is explained and linked to this Kripke model. In this way, it is possible to investigate the influence of the underlying Kripke models to the stories on the accuracy and reaction times of question answering. The story about the chocolate bar and the associated Kripke model are repeated at the end of this section in order to make the reasoning easier to follow.

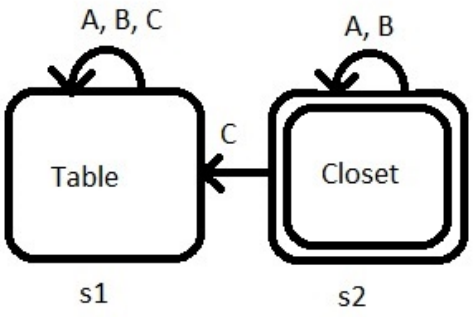
Suppose the question is: "Does A think that B thinks that the chocolate bar is on the table?" To answer this question, you start with what you think. Which is that you are in state s2, where the chocolate is behind the closet. From this state, you follow the accessibility relation for Nina, which goes only from state s2 to itself. This means that you end in state s2, which means that you think Nina thinks the chocolate is behind the closet and therefore not on the table. The question should be answered with "No" and the number of states visited to draw this conclusion is one.

Another example question is: "Does C think that B thinks that the chocolate is on the table?" Again, although this time it is not explicitly stated in the question, you start with reasoning in the state you believe is true: state s2. From this state the accessibility relation for Levi leads you to state s1 and from there the possibility relation for Nina lets you stay in state 1. So you think that Levi thinks that Nina thinks that the chocolate is on the table and therefore the answer to the question should be "Yes". The number of states visited to reach this conclusion is two.

The usefulness of this measure has not yet been proven, but it could be that the transfer from one state to another state increases the reaction time. In particular, this may happen when the questions get more difficult.

Story 1a

Imagine that you (character A) are with Levi (character C) in his living room. Nina (character B) enters with a chocolate bar for Levi, because he had his birthday. Levi puts the chocolate bar on the table. Levi leaves the room to do groceries, Nina stays with you. You decide to hide the chocolate bar behind the closet. Nina sees that you hide the chocolate and you see her surprised look. Then you were tired and went home.



3.3 Summary

In the experiment, six different kinds of questions were used. In Table 1, for each kind of question, an example is given with the associated correct answer and attributes. The story and associated story structure was repeated at the end of last section, so only a quick overview of the question structures in the experiment is presented. For each question structure, an example question, the desired answer and the values of the attributes Order of Knowledge (OoK), Order of Grammar (OoG), Self-Reflexivity (SR), and the number of states (#states) are presented in Table 1.

Structure	Example question	Desired answer	OoK	OoG	SR	#states
1) B	Does Levi think that the chocolate is on the table?	Yes	1	1	No	2
2) AB	Do you think that Nina thinks that the chocolate is on the table?	No	1	2	No	1
3) BA	Does Nina think that you think that the chocolate is behind the closet?	Yes	2	2	Yes	1
4) BC	Does Nina think that Levi thinks that the chocolate is behind the closet?	No	2	2	No	2
5) ABA	Do you think that Levi thinks that you think that the chocolate is on the table?	Yes	2	3	Yes	2
6) ABC	Do you think that Nina thinks that Levi thinks that the chocolate is behind the closet?	No	2	3	No	2

Table 1: question structures

4 Methodologies

In the preceding chapter, the stories and questions for my experiment were discussed. The time participants needed to read and answer the questions was used to answer my research questions. However, I was mostly interested in the time participants needed to formulate their answer, the so-called decision time. Therefore, I needed to distinguish between the reading times and the decision times.

It is not completely certain whether it is legitimate to distinguish between reading times and decision times. Therefore, in this chapter, the method for distinguishing between the two is extensively explained and compared to another well-known method. Also, the experimental results will be analyzed on both the total reaction times and the reading and decision times separately.

In this chapter, first the separation method between reading and decision times is discussed. Then in the second section, linear mixed effects-models are discussed. Finally, some factors that generally influence the reading times are discussed. These factors come helpful forming hypotheses in the next chapter.

4.1 Separating between reading times and decision times

The most obvious and clear method to see experimentally whether there are differences in complexity of the questions would be to analyze the total reaction time needed to answer the questions. However, as the influence of Order of Knowledge and Self-reflexivity are expected to be significantly smaller than the influence of Order of Grammar we expect that their effects might be hard to distinguish. Still, for the completeness we will try to analyze the data.

Another method to measure parsing complexity is a subject-paced reading task (Just, Carpenter, & Woolley, 1982). With this method parts of the questions are sequentially shown and the subject determines when he is finished reading and ready to go to the next part. With this method it can be determined how long the subject needs to read the successive parts of questions. And the decision time can be determined by measuring the time between the last question part was read and the answer was given. However, this method forces the participants to read and evaluate the questions from left to right. Furthermore, they cannot look back at earlier parts of the question. This restriction means that it is impossible to distinguish between question memorization difficulty and question answering difficulty.

Because, the first method is probably not precise enough and the second method asks for too many compromises from the experimental setup, I used a third method. This method uses an eye-tracker to track the participants' eyes. Participants are allowed to

read at their own pace and take all time needed to answer the questions. In this way, the thought processes involved in answering the questions were not interrupted.

The eye-tracker was used to measure the saccades and fixations of the eyes. This information was then used to distinguish between the reading time and the decision time. In the next section, it is explained what saccades and fixations are and how they relate to visual perception in general and reading in particular. Then the algorithm to find the border point between reading, and decision times is discussed.

4.2 Saccades and fixations

People are usually not aware of the effort they unconsciously make to create a continuous image in both space and time. Saccades are small, fast eye movements; the human eye makes around 3-5 saccades per second (Fischer & Weber, 1993). The fact that vision becomes blurred rapidly when the retinal image is prevented from moving (Fischer & Weber, 1993), emphasizes the importance of saccades for continuous viewing.

Saccades are separated by periods of 200-300 ms called fixations (Fischer & Weber, 1993). Even within these fixation periods, the eye slowly drifts back and forth around the mean point of the maintained fixation (Steinman, Haddad, Skavensk, & Wyman, 1973). These eye movements are so small that the subject of the fixation remains in the macula of the retina, where detailed vision is best (Steinman et al., 1973). Consecutive fixations on the same word, while reading, are sometimes aggregated into units called gazes (Just & Carpenter, 1980).

It is possible to track these fixations, saccades, and gazes with an eye-tracker. The eye-tracker used for this kind of analysis (for example, an Eyelink 1000), is a camera with a computer connected to it. With help of this computer both the length and the place of gazes can be determined. In the real-time image analysis the computer measures the eye's pupil and calculates the gazes.

These gazes often correspond to one word (sometimes to a syllable). Therefore, the duration of consecutive gazes corresponds with the time it takes to read a word. Also, it can be detected whether participants reread earlier parts of the question while answering it. In the next section, the assumptions made when distinguishing between the reading times and the decision times are explained.

4.3 Determining reading times

Most people are not aware of the number of eye movements they make when looking at an object (Jacob, 1991). Also, when reading, most people think they are just moving continuously from left to right, sometimes pausing at a word a little longer and in

exceptional cases rereading. When people move their eyes back in text while reading, it is called regression (Rayner, 1998).

It is easy to visualize the eye-tracker data, showing the successive fixations during the recording time. In the experiments participants first read the complete sentence from left to right and then, regressed to earlier information whenever this was necessary. During the initial reading of the question, participants did not regress. The fixation on the last word was chosen as the border point between reading time and decision time.

Eye-tracker data is often imprecise and participant-dependent; therefore, the space between the words of the questions needs to be large enough to make them distinguishable. Also, as the coordinates differed per participant, the measured coordinates of the fixations could not be exactly mapped to the experimental stimuli.

In order to separate between all different words of the questions, the questions needed to be split over two separate lines. These two lines were separated by 48 pixels. Despite of this large space between the two lines, the fixations on those lines could not be separated by the same y-value for all participants.

For each participant the same algorithm was used to segment the fixations into two lines. This algorithm consisted of four steps. First clear outliers were excluded, fixations with a y-coordinate higher than 600 on a screen of 1,024 by 786 pixels. These were fixations on the bottom of the screen and they could only be explained by a participant looking at the keyboard. On the remaining fixations a k-means cluster algorithm (MacQueen, 1967) with two means on the y-coordinates of the fixations was used to split those fixations into two groups. The border between the two groups was estimated to be exactly in between the average y-coordinates of both groups.

The third step of the algorithm was to remove further outliers. This was necessary because a few outliers can shift the border between the two lines and then some fixations would be matched on the wrong line. Fixations that were twice as far from the border between the clusters as the average of the cluster were, therefore, excluded. As a final step, the final border between the clusters was then determined in the same way as that the estimated border was determined, but this time with only the non-outlier subset of fixations.

4.4 Linear mixed-effects models

In statistics, there is a clear distinction between fixed effects and random effects (Baayen, Davidson, & Bates, 2008). Fixed effects typically have a fixed number of levels that can be repeated in time or among participants (Baayen et al., 2008). Fixed effects represent the explanatory variables, such as the number of words in a sentence. Random effects typically originate from individual differences between participants or items. It is

often very difficult to classify these random effects. Mixed effects models try to model both the fixed effects and the random effects. The random effects are modeled with mean zero and an unknown variance. For fixed effects both the mean and the variance are unknown.

The models are fit with a technique called relativized maximum likelihood, also known as restricted, residual or reduced maximum likelihood (Baayen, 2008). This technique is an improvement over the maximum-likelihood technique, as the latter does not consider the loss in degrees of freedom resulting from the estimations (Harville, 1977). This also results in a bias towards a smaller effect (Harville, 1977). Although it is not proven that the relativized maximum likelihood technique has no bias, it is smaller than the bias of maximum likelihood technique.

As the number of levels per fixed effect is limited in the experiments of this thesis, there is little use in analyzing other models than linear models. Therefore, the models used were all linear mixed-effects models.

4.5 Summary

For the analysis of the experiments in this thesis, two main ideas were used. First, the total reaction times to the questions were divided into a reading part and a decision part. This was done using the eye-tracker data.

Furthermore, the statistical analysis was done with linear mixed-effects models, because there was missing data. Furthermore, these models provided a way to evaluate the influence of fixed and random factors at the same time.

5 Research questions and hypothesis

This research was done to learn more about the nature and structure of human reasoning about facts and other people's knowledge. Although it is unknown how knowledge about the world is stored in the human brain, this research aims to provide some restrictions.

As mentioned before (Section 2.1), most researchers investigated whether children in different developmental stages and different kinds of animals were capable of higher-order reasoning about belief and knowledge. In other words whether they possessed Theory of Mind (ToM). In this research the focus is on the speed of the applicability of ToM in human adults.

In the experiments reported in this research, participants needed to read stories and answer the follow-up questions. We manipulated the questions' complexity by introducing structural difference to both the stories and questions (see Chapter 3).

The factors used to measure the complexity of the questions were: Order of Grammar (OoG), Order of Knowledge (OoK), Self-Reflexivity (SR), Story structure, and number of States (see Chapter 3). Those factors were expected to influence the reading time, the decision time or both. In order to distinguish between the reading times and the decision times for the questions, the eye movements of participants were recorded.

5.1 Order of Grammar

The Order of Grammar (OoG) correlates with the length of the question (see Section 3.2.1). Questions with a higher OoG contain more words and are therefore expected to take longer to read. However, there are no indications that the length of the question would influence the reasoning process.

Therefore, the OoG was expected to influence the reading times for questions, but not the decision times. Question with higher OoG were expected to have prolonged reading times than those with lower OoG. In order to specifically test this hypothesis, there were question-pairs that solely differed in OoG, but not in the other factors discussed. An example of such a pair is: "Does Nina think that the chocolate is on the table?" and "Do you think that Nina thinks that the chocolate is on the table".

5.2 Order of Knowledge

The OoK was used to indicate the degree of ToM necessary to answer the question (see Section 3.2.2). In many earlier experiments, this factor has been shown to influence the total reaction times to questions (see Section 2.1). Therefore, this factor is useful to validate the experimental setup of this research.

The influence of OoK on reading times and decision times separately has been investigated less extensively, but for a comparison of reaction times of questions about mental state and non-mental state in autistic individuals see (Bowler, 1997). However, as it reflects reasoning properties, it was expected that OoK will influence the decision times. The decision times for questions with higher OoK were expected to be longer than those of questions with lower OoK. As participants were expected to first read the questions and think of an answer when finished reading, OoK was not expected to influence reading times.

5.3 Self-Reflexivity

In this thesis self-reflexivity is about the necessity to understand another character's perspective on your own knowledge; about the assumptions someone makes about what you, the participant, do and do not know.

Earlier studies have shown that human adults need more time to assess situations where they need to take some else's perspective; and children sometimes seem unable to evaluate such situations (Leslie, 1987). This is especially the case in false belief tasks where they need to reason about someone else's false beliefs (Wimmer & Perner, 1983). Reasoning about someone's false belief about your own beliefs could be seen as the ultimate false belief task. Therefore, questions that required self-reflexive reasoning where expected to be more difficult than other questions. This was expected to result in lower accuracy and longer decision times.

5.4 Story structure

There were three different story structures in terms of Kripke models (see Section 3.1). These Kripke models model the knowledge of the characters at the point where the story was aborted.

Story structure 1 corresponded to a model with only 2 Kripke states, whereas story structure 2 and 3 have 3 Kripke states. Therefore, story structure 1 was expected to be easiest. In structure 2, the participant was aware of the actual state in all Kripke states; but in structure 3, there was a state where the participant did not know the actual state. Therefore, structure 3 was expected to be more difficult than structure 2.

The idea that complex models are more difficult than simpler models is analogous to some work on sentence verification, done by (Szymanik & Zajenkowski, 2010). They showed that natural language quantifiers recognized by finite-automata were easier to understand than those recognized by push-down automata (Szymanik & Zajenkowski, 2010).

5.5 Number of states

Story structures were expected to influence the decision time, as it corresponded to Kripke models that differ in structural complexity. The idea behind this is that the more complex the model is, the more difficult it is for the participant to correctly memorize and use the model. The influence of the story structure on question difficulty may indicate memorization problems. However, it is also interesting to see whether the use of the models influenced the question difficulty.

The use of the models was made explicit by counting the number of unique states that needed to be visited to answer the question (see Section 3.2.4). The hypothesis is that switching states has a higher cost than staying in the same state. This means that the decision times of the question was expected to increase when more states of the Kripke model are needed to be visited to answer the question. The reading times were not expected to be influenced by this factor.

6 Experiment 1

The goal of this experiment was to find the factors that influence the total reaction times, the reading times, and the decision times. The total reaction times were expected to be increased for higher Order of Grammar (see Section 5.1) and higher Order of Knowledge (see Section 5.2). The reading times were expected to be increased by only higher Order of Grammar. The decision times were expected to be increased by higher Order of Knowledge and self-reflexivity (see Section5.3).

6.1 Methods

6.1.1 Subjects

In total, 20 students (12 males and 8 females) of the Hanze University and the University of Groningen were selected for the experiment. Participants had normal or corrected-to-normal (by lenses) visual acuity, their ages ranged from 19 to 28 years. All participants were offered a free drink in return for their participation. Informed consent was obtained from each participant.

6.1.2 Apparatus & Materials

Both the stories and questions (explained in detail in Chapter 3) were presented in black on a white background in a bold 18-point Courier New font. They were presented in the middle of a 20-inch computer screen set at a resolution of 1,024 by 786 pixels. Viewing distance was about 60 cm. An EYELINK 1000 eye-tracker was used to record the eye movements of the dominant eye, at a sample rate of 500 Hz. The task information was displayed on a computer in an experiment constructed with E-prime (http://www.pstnet.com/eprime.cfm). This program supports the use of an eye-tracker and can also measure reaction times.

Only 80% of the width of the screen was used. The sides were black to prevent participants from looking at that part of the screen. This is because the eye-tracker may be less accurate there. Because of the use of word-wrap, approximately 80% of the frame was used for the actual story presentation. The questions were all divided over two lines; on the first line the part like “Do you think that Levi thinks that ” and on the second line, the part like “the chocolate bar is on the table?” was presented. There were two empty lines between the two lines. The first line used approximately 80% of the frame; the second line took up to approximately 80% of the frame, depending on the Order of Grammar.

For the analysis of the results of this experiment statistics R and Excel were used, along with the LME4-package (Bates, Maechler, & Dai, 2008).

6.1.3 Procedure

The experiment lasted about half an hour for each participant. First the dominant eye was determined. The eye-tracker was calibrated in about 5 minutes. Then the actual experiment was started. The task-specific instructions were read by the participant from the computer screen. The subjects' eye fixations and movements were recorded only while participants were answering the questions.

The experiment stories discussed in Section 3.1 were presented one by one, followed by questions. Each story was presented as a whole. Participants were asked to memorize the story and press the space bar after they read, understood, and memorized the story. Then the story disappeared and the first question appeared. After answering the question, it disappeared, and the next question appeared until all the questions for the story were answered.

Every participant needed to answer exactly one question per condition for each story. The questions were grouped on the OoG, first the 1st-order-question, then the three 2nd-order-questions, and finally the two 3rd-order-questions. The order of the questions within each OoG was random.

The first practice story with its related questions was the same for each participant and was excluded from the analysis. Participants knew it was a practice story and were encouraged to ask remaining questions during this practice session.

To prevent unexpected and unwanted performance according to story order, the order of the rest of the stories was randomized over participants. Furthermore, the experiment was composed in such a way that the numbers of “yes”-and “no”-questions were equal. On overview of all the stories may be found in Appendix A. All possible questions for the example story may be found in Appendix B.

6.1.4 Analysis

Eye fixations were measured during question-answering. The eye-tracker data was used to determine the total reaction times, the reading times, and the decision times (see Section 4.3).

For the statistics, linear mixed effects models (LME models) were used (see Section 4.4). As the recorded measures were all right-skewed, the analysis was performed on the logarithm of the measures. As individuals typically show different reaction times, Subject was chosen as a random effect. Also the stories differ in difficulty, so the story was another random effect.

The total reaction, reading, and decision times were fitted with LME models using REML technique (Gurka, 2006). The models were fit on the logarithm of these variables, as they were left-skewed. For all models, subject and story were chosen as

random factors. The significance level was 0.05. The full model is represented in Model 1. It includes the basic fixed effects that were compared for the total reaction, the reading, and the decision times.

$$y = \mu + \beta_1 OoG + \beta_2 OoK + \beta_3 self\text{-}reflexivity + \beta_4 number\ of\ states + \beta_5 story\ structure + Error(Subject) + Error(Story)$$

Model 1: full model with all fixed effects

The order of grammar (OoG) reflects the length of the question (see Section 3.2.1). The order of knowledge (OoK) reflects the degree of theory of mind needed to answer the question (see Section 3.2.2). The self-reflexivity reflects whether there were conflicting models of oneself necessary to answer the question (see Section 3.2.3). The story structure corresponds to the structure of the story the question was about (see Section 3.1). The fixed effect number of states reflects the number of states needed to visit in the Kripke model to answer the question (see Section 3.2.4).

To determine the best model to fit the data, first the full model was analyzed. This model contained all fixed factors (see Model 1). One by one, these factors were excluded and then it was analyzed whether the simpler model performed significantly worse than the earlier model. If so, the simpler model lost explanation power and probably is worse.

However, for marginal cases, the Akaike information criterion (AIC) was used (Akaike, 1974). The model with the lowest AIC probably was the best model. The AIC considers both the complexity of the model and the goodness of fit.

6.2 Results

In this section, the total reaction times, the reading times, and the decision times for the questions to the stories were analyzed. The goal was to find the models that best predict these times. The results were summarized within one graph, to provide some understanding of what the results are about (Figure 4).

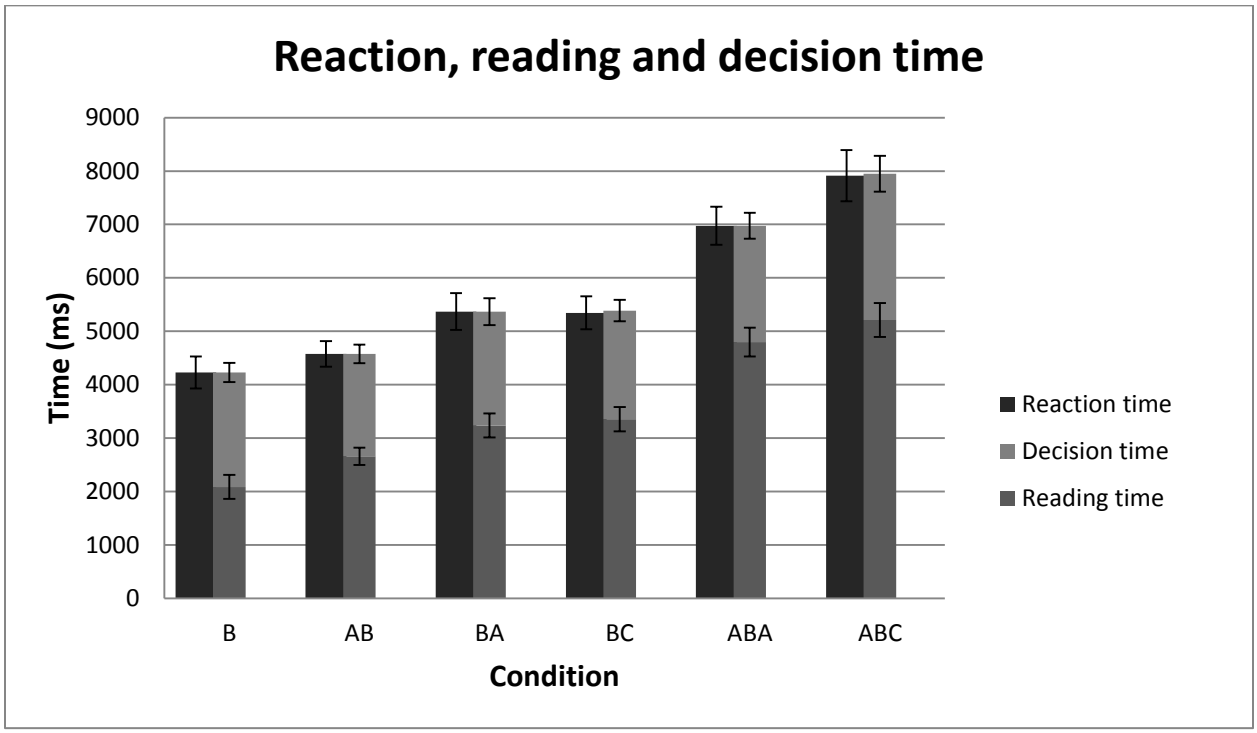


Figure 4: Summary of the results of Experiment 1

In Figure 4, the total reaction times, the decision times, and the reading times were shown per condition. The condition refers to the kind of question that was answered. The letters in the condition correspond with the characters in the question. Character A was the participant himself and the characters B and C could either be Nina or Levi (see 3.2 for a more complete explanation). The reading times and the decision times add up to the total reaction times. There are little discrepancies because for questions of some participants, the border between the reading and decision time could not be found due to missing eye-tracker data.

In Figure 4, it can also be seen that the reading time part of the reaction time is larger than the decision time part. Therefore it may be difficult to correctly distinguish the effects of the decision time. Also, the total reaction times seem to increase with question length, as does the reading time. However, the decision times seem to be almost constant among the conditions. In the next few sections, these effects were analyzed extensively and explained.

6.2.1 Total reaction times

Starting with Model 1, the least contributing factors for the total reaction times were the story structure (n.s., $\chi^2 = 1.5629$) and the self-reflexivity (n.s., $\chi^2 = 0.6494$). Dropping the next least contributing factor, the number of states, did decrease the performance of the model (p = 0.04003, $\chi^2 = 6.436$). This meant that dropping the number of states factor decreases the performance of the model.

$$y = \mu + \beta_1 OoG + \beta_2 OoK + \beta_4 number\ of\ states + Error(Subject) + Error(Story)$$

Model 2: possible model for total reaction time

$$y = \mu + \beta_1 OoG + \beta_2 OoK + Error(Subject) + Error(Story)$$

Model 3: possible model for total reaction time

However, considering the complexity of the models, I noticed that the extra explanation power of Model 2 (AIC = 1414) does not equipoise for the extra complexity when compared with Model 3 (AIC = 1408). Therefore, the best possible model to predict the total reaction time is Model 3, with both the OoG and the OoK as a fixed effect.

In Table 2, the statistics of the Model 3 are listed; the factors OoG is 1 and OoK is 1 were used as a baseline and were reflected in the intercept.

Factor	Estimate	Std. Error	t-value
(intercept)	8.21563	0.08996	91.32
OoG is 2	0.11286	0.05339	2.11
OoG is 3	0.44576	0.06539	6.82
OoK is 2	0.12083	0.04624	2.61

Table 2: factors of OoG + OoK model for total reaction times

However, as expected before, the reaction time consists of two different components: the reading time and the decision time. To fully understand why both higher OoG and higher OoK increased the reaction time, the same LME models should be applied to these sub-parts of the reaction time.

6.2.2 Reading times

For the reading times, the least contributing factors were story structure (n.s., $\chi^2 = 3.1024$), self-reflexivity (n.s., $\chi^2 = 1.5056$), and number of states (n.s., $\chi^2 = 4.4823$). The next least contributing factor, the OoK did decrease the model's performance (p = 0.00121, $\chi^2 = 10.475$). Considering the model complexity, the model with OoK ($AIC = 1664$) should still be preferred over the simpler model without OoK ($AIC = 1668$).

$$y = \mu + \beta_1 OoG + \beta_2 OoK + Error(Subject) + Error(Story)$$

Model 4: best fitting model for the reading times

In Table 3, the estimated influence calculated by the model was shown for all fixed effects. The t-values of these factors show that all these effects were significant.

Factor	Estimate	Std. Error	t-value
(intercept)	7.46367	0.09314	80.13
OoG is 2	0.30684	0.06341	4.84
OoG is 3	0.72582	0.07766	9.35
OoK is 2	0.17784	0.05488	3.24

Table 3 factors of OoG + OoK model for reading times

The results for the reading times were not expected. The hypothesis was that only higher OoG would increase the reading times, instead higher OoK also increased the reading times. This may be explained by participants starting to think while reading. However, I noticed that questions with OoK 2 had on average more character names ("Nina" and "Levi") than OoK 1, for the same OoG-levels. It was expected that these character names would take longer to read than a simple "you".

Adding the number of names (number of times "Nina" or "Levi" appears in the question) as a fixed effect to Model 4, does not improve the model (n.s., $\chi^2 = 2.757$). Also, the alternative model, Model 5 (AIC = 1667), does not outperform Model 4 (AIC = 1664).

$$y = \mu + \beta_1 OoG + \beta_2 OoK + Error(Subject) + Error(Story)$$

Model 5: alternative model for reading times

6.2.3 Decision times

For the decision times, the least contributing factors were OoK (n.s., $\chi^2 = 0.0053$), self-reflexivity (n.s., $\chi^2 = 0.2363$), story structure (n.s., $\chi^2 = 1.0485$), and number of states (n.s., $\chi^2 = 3.8122$). Dropping the last factor, OoG, did decrease the model performance (p = 0.007464, $\chi^2 = 9.7952$).

However, considering the complexity of the models, the model with no fixed effects ($AIC = 2552$) should be preferred over the model with OoG as a fixed effect ($AIC = 2553$). Because the difference in AIC was small, the conclusion that the model with no fixed effects was better than the model with OoG as a fixed effect cannot conclusively be drawn. But also, the influence of the OoG on the decision times was not demonstrated.

6.3 Discussion

The primary goal of this experiment was to better understand the underlying structures of ToM in adults. The total reaction times and reading times were shown to be prolonged for higher OoG. These results were to be expected. It is obvious that longer questions take longer to read.

Furthermore, none of the fixed effects seemed to influence the decision times. This was not expected and indicated that the questions in the experiments were too easy. Therefore, the decision times were so small that differences could not be distinguished.

It was a little surprising that the reading times increased for questions with higher OoK. A possible explanation was the number of times “Nina” and “Levi” appeared in the questions in the different OoG-levels. This was higher for OoK 2, than for OoK 1. As less frequent words generally take longer to read than more frequent words (Rayner & Duffy, 1986), questions about a “Nina” or a “Levi” character will take a little bit longer to read than questions about the “you”-character. This was evaluated and the model with the number of names instead of OoK for the reading times did not outperform the model with OoK. So this effect was not shown in this experiment.

This finding, in combination with the fact that no fixed effects were found the influence the decision times makes another explanation more likely. Participants were thinking of the answer while they were reading the question.

OoG and OoK are clearly not independent. Therefore, it might seem to make sense to look at the interaction effects of those two variables. However, this was impossible as for OoG is one, only OoK one is possible, and for OoG three, only OoK two was used. This last fact was a choice to keep the number of questions per story low. Otherwise, it might be that participants needed to rethink about the story during the question answering.

There were many participants who mentioned the characters Nina and Levi after the experiments. Those two characters were randomly mapped on the roles in the story. However, participants attributed several characteristics to them such as: smart, dumb, and sweet. When I explained that their roles were random, they were a little disappointed.

However, these conversations indicated that the idea to use just two foreign character names had worked at least partly. The idea was that the participants were so familiar with the characters, that the reading times of the character names would decrease and also that they would be less informed about the goal of the experiment.

Also, it was found that the decision time on the first question was largest of all questions. The first question (condition B) had the form: “Does X think that …?”. This is contra-intuitive as it was supposed to be the easiest question. Probably, participants needed time to recollect their memory of the story before they answer the question. Somehow, this special memory recollection was stored afterwards in the short-time memory circuit. Participants also indicated that they needed to recollect their memory of the story after reading it, especially to remember who is who in the story.

Although the main purpose of this experiment was to find the influence of self-reflexivity, it proved to be impossible in this design. The differences in decision times were so small, that they could barely be distinguished. Therefore, in a sequel experiment, these decision times should be larger.

To increase the decision times, there are roughly two possibilities: increasing the question difficulty and increasing the mental workload during question answering. As increasing the question difficulty would lead to longer questions and more logically formulated questions, the participants would be more likely to discover the experimental setup during the experiment.

Also, increasing the question difficulty does not prevent participants to start thinking while they are still reading. In order to make a clearer distinction between reading and deciding, the best option was to increase the work load. This was done in the next experiment, by asking participants to do another task, parallel to the question-answering. This task required working memory and therefore increased the work-load.

7 Experiment 2

The results of Experiment 1 showed that participants probably started thinking while still reading the questions. This caused difficulties in identifying the factors influencing the decision times. Therefore, in this experiment, the experimental load was increased.

This mental load was increased by blocking part of the working memory of the participants. This was achieved by asking participants to remember 3-digit numbers while answering the question. Apart from this side-task, the experimental setup was kept the same.

The goal of this experiment was also the same as that of Experiment 1: finding the factors that influence the total reaction times, the reading times, and the decision times. The total reaction times were expected to be increased for higher Order of Grammar (see Section 5.1) and higher Order of Knowledge (see Section 5.2). The reading times were expected to be increased for by only higher Order of Grammar. The decision times were expected to be increased for higher Order of Knowledge and self-reflexivity (see Section5.3). The influence of self-reflexivity on the decision times was expected to be too small to be significantly distinguished in the total reaction times.

7.1 Methods

7.1.1 Subjects

In total, 22 students and former students (16 males and 6 females) of the Hanze University and the University of Groningen were selected for the experiment. Participants had normal or corrected-to-normal (by lenses) visual acuity, their ages ranged from 20 to 31 years. All participants were offered a free drink in return for their participation. Informed consent was obtained from each participant.

7.1.2 Materials

For this experiment, the same stories and questions were used as for Experiment 1 (explained in Chapter 3). Furthermore, there was a memory game, the 'higher/lower game'. For this game, participants needed to remember a 3-digit number. This number was compared to another 3-digit number that was shown next. This new number then needed to be remembered and compared to the next number. This means that each number was compared to the preceding number. When the participant did not answer correctly, a red screen was shown. The game was restarted by showing a fresh number after reading the story.

7.1.3 Apparatus

All stimuli were presented in black on a white background in a bold 18-point Courier New font in the middle of a 20-inch computer screen set at a resolution of 1,024 by 786

pixels. Viewing distance was about 60 cm. Participants used a US-English keyboard to interact with the experiment.

Only 80% of the width of the screen was used. The sides were black to prevent participants from looking at that part of the screen. This is because the eye-tracker may be less accurate there. Because of the use of word-wrap, approximately 80% of the frame was used for the actual story presentation. Also the questions were presented in two lines, so they also used approximately 80% of the frame. The 3-digit number was shown in the middle of the screen. The red screen shown for a wrong answer to the higher/lower-game used 100% of the screen.

An Eyelink 1000 eye-tracker was used to record the eye movements of the dominant eye, at a sample rate of 500 Hz. Recordings started as soon a question was represented on the screen and ended when the participant had answered the question. For the analysis of the results of this experiment statistics R and Excel were used, along with the LME4-package (Bates et al., 2008).

7.1.4 Procedure

First the eye-tracker was calibrated to the dominant eye. Then the participant read the instructions from the screen. The participant could ask questions during the first practice block. This block was not included in the analysis. It started with the presentation of the story. The participant could take as long as needed to read and memorize the story. After pressing the space bar, the first number of the higher/lower game appeared. After the participant memorized this number, she could press any key to go on.

At first B-condition question was presented. The participant could answer this question by pressing ‘j’ for yes and ‘k’ for no. After answering this question, the second number of the higher/lower game appeared. Participants were asked to press ‘e’ when the number was higher than the preceding number, and to press ‘d’ when the number was lower than the preceding number.

After answering to the higher/lower game, the next question appeared and after each question the higher/lower game appeared again, until all six questions were asked. The questions of conditions AB, BA and BC appeared in random order first. After these questions, the questions of condition ABA and ABC appeared in random order. So the questions were ordered with respect to the OoG and randomized within OoG.

After the practice block, the participant was informed that the real experiment was about to start and that any questions could be asked at that time. The experiment consisted of eight stories along with their corresponding questions. The procedure was exactly the same as for the practice block, but now repeated eight times to cover all

stories. At the end of the experiment, participants were asked to call for the experimenter and were offered a cup of coffee.

7.1.5 Analysis

Three different variables were analyzed for this experiment: the total reaction times, the reading times, and the decision times. These were determined with the eye-tracker data (see Section 4.3).

The reaction, reading, and decision times were fitted with LME models using the REML technique (Gurka, 2006). Subject and story were chosen as random factors for all models used. The explanation for this choice can be found in Section 6.2. To determine whether two models differed from each other, ANOVAs were used with a 0.05 significance level.

At first, the model with all thinkable variables was made. After this, recursively, the least contributing variable was excluded and tested against the preceding model. This continued until the calculated model differed significantly from the preceding model. Then the simplest model was found that still explains the data. The Akaike information criterion (AIC) was used (Akaike, 1974) to balance between model complexity and explanation power. The model with the lowest AIC probably was the best model. The AIC considers both the complexity of the model and the goodness of fit.

As the total reaction time, the reading time, and the decision time were left-skewed, the models were fitted on the logarithm of these values. The models consist of combinations of the fixed factors OoG, OoK, self-reflexivity, number of states, and story structure. They have Subject and Story as random effects. The outcome variable is represented by 'y', this value is the prediction of the dependent variables.

$$y = \mu + \beta_1 OoG + \beta_2 OoK + \beta_3 SR + \beta_4 nStates + \beta_5 storyStructure + Error(Subject) + Error(Story)$$

Model 6: possible fixed effects for the total reaction times, reading times, and decision times

The order of grammar (OoG) reflects the length of the question (see Section 5.1). The order of knowledge (OoK, see Section 5.2) reflects the degree of theory of mind (ToM) needed to answer the question. The self-reflexivity reflects whether there were conflicting models of oneself necessary to answer the question (see Section 5.3). The story structure corresponds to structure of the story the question was about (see Section 5.4). The fixed effect number of states reflects the number of states needed to visit in the Kripke model to answer the question (see Section 5.5). So it does not depend on the complexity of the Kripke model of the story, but only on the complexity of the part of the model that needed to be visited.

For one participant, the eye-tracking did not work properly. Therefore, his data could only be used for the analysis of the total reaction times, but not for the reading times and the decision times.

7.2 Results

In this section, the goal was to find the best fitting model for the total reaction times, the reading times, and the decision times. The expectations were the same as they were originally for Experiment 1. The total reaction times were expected to be increased for higher Order of Grammar and higher Order of Knowledge; the reading times by higher Order of Grammar; and the decision times by higher Order of Knowledge and the self-reflexivity.

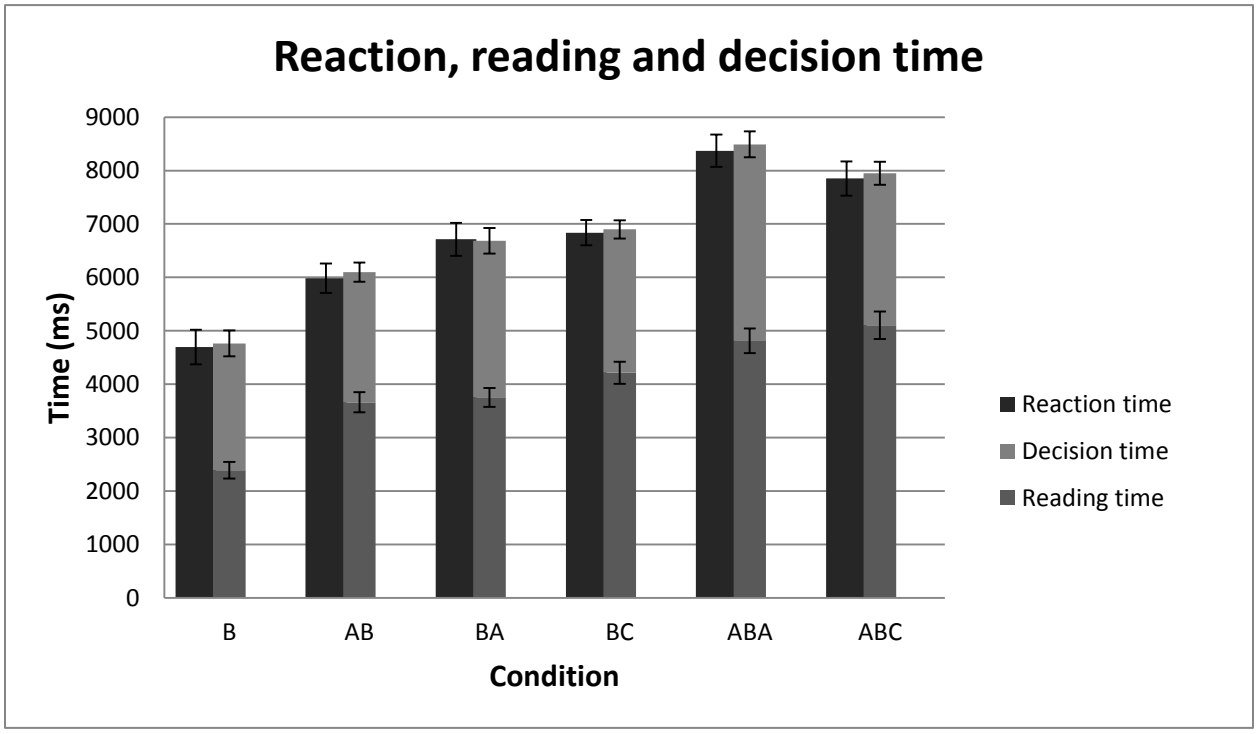


Figure 5: summary of the results of Experiment 2

In Figure 5, the total reaction times, the decision times, and the reading times were shown per condition. The condition refers to the kind of question that was answered. The letters in the condition correspond with the characters in the question. Character A was the participant himself and the characters B and C could either be Nina or Levi (see Section 3.2 for a more complete explanation).

Normally, the reading times and the decision times would add up to the total reaction times. However, for one participant no eye-tracker data was collected. Also, for some other participants, the border between reading and decision times could not conclusively be determined. Therefore, the graph shows some discrepancies.

In Figure 5, it can be seen that the reading time seems to be larger than the decision part and that it increases when the Order of Grammar increases. For the decision times, it is more difficult to see patterns. Therefore, in the next few sections, the total reaction times, the reading times, and the decision times are discussed separately.

7.2.1 Total reaction times

To find the best predicting model for the total reaction times, Model 6 was simplified step by step. The following variables could be excluded without decreasing the performance of the model: number of states (n.s., $\chi^2 = 0.0126$), self-reflexivity (n.s., $\chi^2 = 0.0981$), and story structure (n.s., $\chi^2 = 3.3151$). When the OoK was excluded, the model did perform less (p = 0.0002473, $\chi^2 = 13.432$).

$$y = \mu + \beta_1 OoG + \beta_2 OoK + Error(Subject) + Error(Story)$$

Model 7: Best fitting model for total reaction times

When the complexity of the models was also evaluated with the AIC, Model 7 (AIC = 1375) seemed more fitting, than the simpler model with only OoG (AIC = 1382). This means that the model with just OoG and OoK is best to explain the total reaction times.

However, the interaction model of OoG and self-reflexivity (see Model 8) performed better than the simpler model with just the fixed effects of OoG and self-reflexivity (p = 0.02821, $\chi^2 = 4.8152$). However, considering the complexity of the models, Model 8 ($AIC =$ 883.1) was worse than Model 7 ($AIC = 882.3$). Therefore, this more complex model was rejected.

$$y = \mu + \beta_1 OoG * self - reflexivity + Error(Subject) + Error(Story)$$

Model 8: possible interaction effects for total reaction times

So the best predicting model for the total reaction times had both OoG and OoK as fixed effects. A summary of this model is given in Table 4. In this summary, OoG is 2 and OoK is 1 were used as a baseline reflected in the intercept.

Factor	Estimate	Std. Error	t-value
(Intercept)	8.54721	0.06467	132.17
OoG is 1	-0.26871	0.04726	-5.69
OoG is 3	0.19126	0.03342	5.72
OoK is 2	0.15027	0.04093	3.67

Table 4: OoG/OoK model for the total reaction time

The results show that all levels of the fixed effects had a significant effect on the model; this was reflected in the t-value. They show that the total reaction times increase both when the OoG increases and when the OoK increases. The effect of the OoK seems to be

smaller than that of the OoG, which is not surprising, as the reading times were larger than the decision times.

The effects of OoG and OoK were to be expected and are consistent with the results of experiment 1. However, the hypothesis is that the reading times and the decision times were influenced by different factors. Therefore, in the next two sections, models with these factors were also fit to the reading times and the decision times.

7.2.2 Reading time

Starting with the same model as for the total reaction times (Model 6), the following effects were excluded in order of mentioning: number of states (n.s., $\chi^2 = 0.4708$), story structure (n.s., $\chi^2 = 0.1214$), self-reflexivity (n.s., $\chi^2 = 1.9751$), and OoK (n.s., $\chi^2 = 1.9426$). Dropping the last fixed effect OoG did decrease the performance of the model ($p < 2.2 \cdot e^{-16}, \chi^2 = 157.3$). Considering both complexity and performance showed that the OoG model ($AIC = 1947$) indeed outperformed the model without any fixed effects ($AIC = 2091$).

Factor	Estimate	Std. Error	t-value
(Intercept)	8.05954	0.07507	107.36
OoG is 1	-0.53758	0.05629	5.45
OoG is 3	0.24148	0.04430	13.06

Table 5: OoG model for the reading times

Table 5 shows the estimates, standard errors and t-values for the fixed factor-levels OoG with OoG is 2 as a baseline reflected in the intercept. The results show that both the differences between OoG 1 and 2 and the differences between OoG 2 and 3 can be considered significant. Furthermore, they show that the reading times increase when the OoG increases.

It is interesting that the difference between the first two levels seems to be larger than the difference between the last two levels even though the difference in the number of words added is minimal. A possible explanation was provided by the participants. They indicated that they often skipped the “Do you think”-part of the question as this does not change the outcome of the question. This beginning appears in every question with OoG is 3 and only in one third of the questions with OoG is 1. Therefore, the reduction in reading times was bigger for OoG is 1 than it was for OoG is 2.

7.2.3 Decision time

Effects were excluded from the full model, in the following order: OoK (n.s., $\chi^2 = 0.0228$), number of states (n.s., $\chi^2 = 0.6073$), and story structure (n.s., $\chi^2 = 5.1478$). Excluding the next least contributing effect – OoG, decreased the performance of the

model ($p = 0.02712$, $\chi^2 = 7.2151$). So the best performing model contains the following effects: OoG and self-reflexivity.

Considering the complexity of the model as well, shows that the model with OoG and self-reflexivity ($\mathrm{AIC} = 2495$) did not outperform the model with only self-reflexivity ($AIC = 2491$). So the best possible model considering both performance and model complexity seems to be the model with just self-reflexivity as a fixed effect.

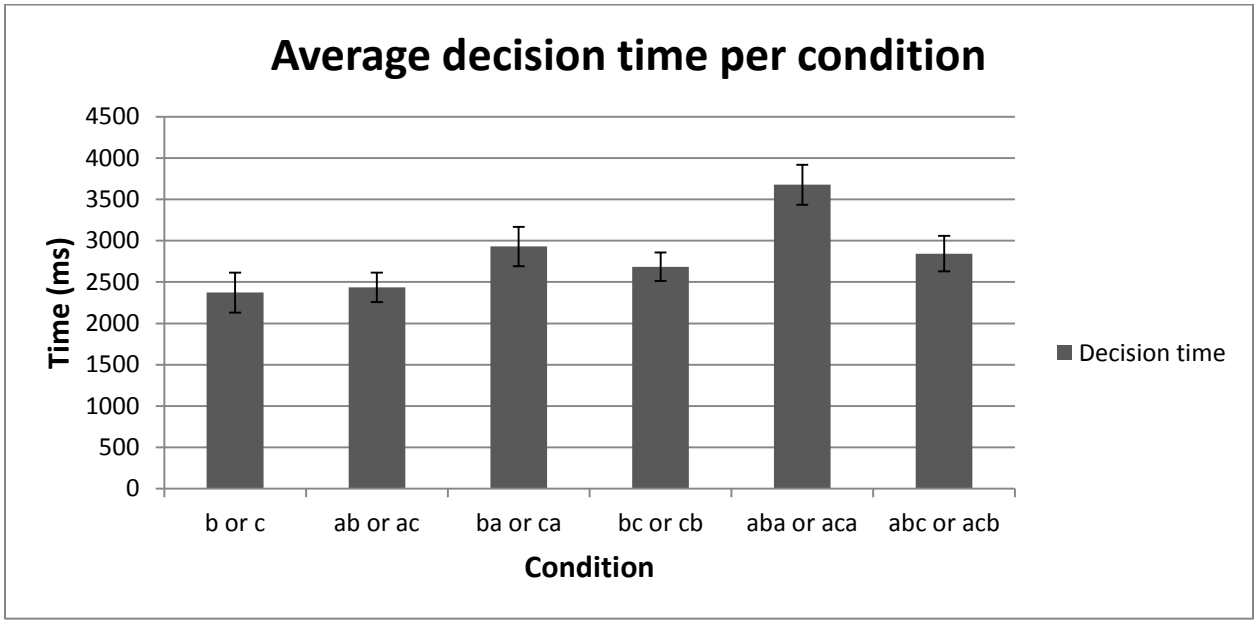


Figure 6: decision times

In Figure 6, the effect of self-reflexivity on condition BA seems smaller than for condition ABA. Therefore, it is interesting to investigate the interaction effects of OoG and SR. However, the performance of the model with the interaction effect was not significantly better than the performance of the model without the interaction effect (n.s., $\chi^2 = 3.7948$).

Possible interaction effects of story structure with self-reflexivity cannot be seen in the graph, but are imaginable because the story structure might influence how the story information was structured in the brain. However, adding this interaction effect did not increase the model performance (n.s., $\chi^2 = 3.1326$).

Considering the complexity of the models as well as some possible interaction effects, the best model for the decision times consisted of solely the self-reflexivity. A summary of this model was given in Table 6. These results show that when a question was self-reflexive, then the decision times increased. In contrast, the OoK, which was also expected to influence the decision time, was the first factor to be dropped.

Factor	Estimate	Stad Error	t-value
(Intercept)	7.53633	0.11423	65.97
SR is Yes	0.13247	0.06636	2.00

Table 6: Self-reflexivity model for decision times

7.3 Discussion

HIgher Order of Knowledge (OoK) and higher Order of Grammar (OoG) were shown to increase the total reaction times. This is consistent with the finding that the reading time depends mainly on the OoG. However, the decision time mainly depended on the self-reflexivity and not at all on the OoK.

An interesting question is why the total reaction times were significantly increased for higher OoK, but neither the reading times nor the decision times seemed to be influenced by this factor. However, combining the results for the reading times and the decision times provides an explanation on how this is possible.

The questions with an OoK of 2 either have an increased decision time caused by self-reflexivity (conditions BA and ABA), or they have an increased reading time caused by the number of 4-letter names in the question (conditions BC and ABC). The questions with an OoK of 1 are neither self-reflexive or have a higher number of 4-letter names in them (conditions B and AB).

These results suggest that the OoK may not cause the increase in difficulty of the questions, but may merely be a summation of two other effects: the effect of having foreign names in the questions increasing the reading times and the effect of self-reflexivity.

At last, there was some problem with the experimental setup. Story 8 (see Appendix A – Stories) asked participants to remember at what time the tigers were fed. This did not cause problems in Experiment 1. However, in this experiment participants also needed to remember numbers during question-answering for the higher/lower-game. Therefore, the performance of the participants on this story was lower than average in this experiment.

Also, among the participants was one person who did not accept the questions to the story. He thought it was impossible to answer the questions as there would be no way that the characters in the story would not suspect they were fooled. He expected them to be suspicious.

8 General discussion

In this thesis, I tried to investigate the everyday reasoning of human adults. This was done by reading experiments with stories and questions to these stories. In Chapter 6, the first experiment was discussed. The stories were constructed in such a way that the agents' beliefs at the end of the stories could be represented with one out of three possible types of Kripke models that we delineated. The questions to those stories all matched the following pattern: Does X believe (that Y believes) (that Z believes) that …?

The total reaction times, the reading times, and the discussion times were used to measure the difficulty of the questions. An eye-tracker was used to separate reading times from decision times. Possibly influencing factors to the question difficulty were: the length of the question, the Order of Knowledge necessary to answer the question, whether the question was self-reflexive (asked about another character's belief about the belief of the participant), and the number of states in the Kripke model that needed to be visited in order to answer the question. Furthermore, the story structure was considered a possibly influencing factor on the question difficulty (see Chapter 5).

Experiment 1 (see Chapter 6), just required participants to read the stories and answer the corresponding questions. The results of this experiment showed that the total reaction times and the reading times were increased by the same factors: length of the questions and order of knowledge. On the other hand, the decision times did not seem to be influenced by anything. The decision times were also short (around 1.5 s) and they did not really vary among the question structures. Furthermore, participants did not reread parts of the questions during the decision times. Therefore, I suspect that the participants were combining reading and thinking.

In Experiment 2 (see Chapter 7) the story reading task was extended with a dual-task: a higher/lower game. This task was meant to occupy some of the working memory of the participants in order to prevent the participant from thinking while reading. This seemed to work, as in this experiment, the reading times were solely increased by the length of the questions and they were not influenced by any of the other factors. The factors increasing the total reaction times were the length of the question and the order of knowledge, consistent with the first experiment. The decision times were shown to be increased for self-reflexive questions.

The finding that the total reaction times were shown to increase with the question length and order of knowledge in both experiments was consistent with the hypothesis (see Sections 5.1 and 5.2). The ability to pass a second-order Theory of Mind – task is acquired very early, around the age of 6 (Leslie, 1987). Therefore, there has not been that much research on the ability of adults to perform theory of mind such as second-

order false-belief tasks. However, even in adults, theory of mind is not always reliably used (Keysar, Lin, & Barr, 2003).

One common explanation is that the knowledge of adults about a situation interferes with the ability to reason about other people's beliefs about the situation. This idea has been called the curse-of-knowledge (Birch & Bloom, 2007). It has been tested in this thesis by comparing self-reflexive questions with non-self-reflexive questions. An example of a self-reflexive question is: "Does Nina think that *you* think that the chocolate bar is on the table?". The non-self-reflexive counterpart is: "Does Nina think that *Levi* thinks that the chocolate bar is on the table?". As the participant is certain about her or his beliefs, this knowledge was expected to interfere with the actual question about what the other character believes about the participant's belief.

The self-reflexive questions were significantly shown to require on average longer decision times than non-self-reflexive questions in Experiment 2. Actually, this factor was the only factor influencing the decision times in Experiment 2. The Order of Knowledge was not shown to influence the decision times at all.

However, the Order of Knowledge was shown to influence the total reaction times, so it was expected to influence either the reading times or the decision times. To understand how this is possible, we need to reassess the reading times of the questions. Longer questions were expected to have longer reading times and therefore also longer total reaction times. The length of the question was not expected to influence the decision times. This hypothesis was confirmed in both experiments reported in this thesis.

Another factor that influenced the reading times was the word choice in the questions. Less frequent words take more time to read than more frequent words (Rayner & Duffy, 1986). The word choice among the questions was kept as constant as possible and was otherwise balanced. For example, for the middle length: 'Does X think that Y thinks that the chocolate bar was on the table?' And 'Does X think that Y thinks that the chocolate bar is behind the closet?' The endings of these questions were balanced within stories and between participants and were therefore not expected to influence the reading times. However, the characters could either have a 4-letter name, namely: "Nina" or "Levi" or there could be a reference to the participant: "you". As the first character option is less frequent, it takes longer to read. In Experiment 2 (see Chapter 7), the reading times were found to depend solely on the length of the question. The longer the question, the longer it took to read. Also, questions with two 4-letter names had longer reading times than questions with only one 4-letter name. Although this effect was not significant in the analysis of the reading times, it does offer an explanation for the fact that questions with a higher Order of Knowledge increase the total reaction times, but influence neither the reading times nor the decision times.

This effect also provides an explanation for the finding that self-reflexive questions have prolonged decision times, but no prolonged total reaction times. This is because the group of self-reflexive questions combined with the group of questions with two 4-letter names (instead of one 4-letter name) form exactly the group of questions with an Order of Knowledge of 2. So every question with an Order of Knowledge of 2 either has a prolonged reading time caused by the number of 4-letter names in the question, or a prolonged decision time caused by the self-reflexivity of the question.

This explanation seems to be in conflict with the results of Experiment 1. In this experiment, the reading times did not seem to have a prolonged reading time for questions like: “Does Nina think that Levi thinks that …?” when compared to questions like: “Does Nina think that you think that …?”. Especially since both these questions have long reading times when compared to questions like: “Do you think that Nina thinks that …?”. An explanation for this finding was provided by the participants. They indicated that they tended to skip the “Do you think”-part of the question as soon as they recognized it, because they knew that it did not impact the answer of the question. This behavior was not reported by participants in Experiment 2.

9 Future work

In this thesis, every day Theory of Mind (ToM) reasoning was investigated. This was done by reading experiments: participants needed to read stories and answer the corresponding questions. The stories were constructed in such a way that the agents' beliefs at the end of the stories could be represented with one out of three possible types of Kripke models that we delineated. The questions to those stories all matched the following pattern: Does X believe (that Y believes) (that Z believes) that …?

The influence of the following factors on question difficulty was investigated: length of the question (see Section 3.2.1), Order of Knowledge (see Section 3.2.2), Self-reflexivity (see Section 3.2.3), number of states (see Section 3.2.4), and story structure (see Section 3.1). The question difficulty was measured by the total reaction times, the reading times, and the decision times (see Chapter 5). The total number of eye fixations and the number of regressions were also candidates as a measure for question difficulty.

Self-reflexivity (SR) in this thesis is used for situations where a character needs to make inferences about what someone else believes about his/her own beliefs. In contrast to inferences made about the beliefs of someone else about facts or other characters' beliefs.

9.1 Story structures

In the experiments of this thesis, three different story structures were used. In both experiments, the story structures were not shown to influence the total reaction times, the reading times, or the decision times. In Experiment 1, the results did suggest an interaction effect between story structure and self-reflexivity on the total reaction times (see Figure 1). The figure suggests that questions that require self-reflexivity clearly increase the total reaction times for story structures 1 and 2 but not for structure 3. This suggests that the influence of a question requiring self-reflexivity is larger for simpler structures. Also, in this figure, there is no suggestion of influence of the story structure on the total reaction times.

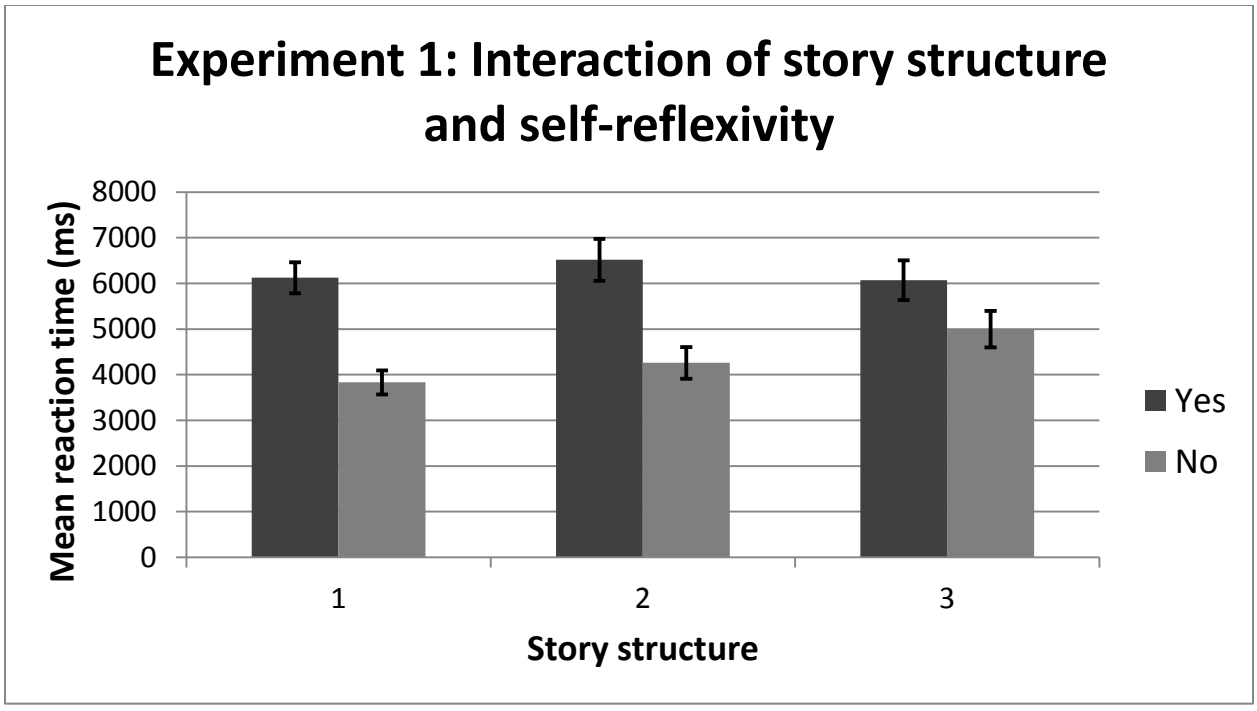


Figure 7: the influence of story structure and self-reflexivity on the total reaction times in Experiment 1

In Figure 8, the analogous results are shown for experiment 2. In this graph no interaction effects of self-reflexivity are suggested, the influence of self-reflexivity on the total reaction times does not seem to depend on story structure. However, story structure 2 seems to be a little bit more difficult than the other two structures. But these results may just originate from the specific, non-structural aspects of the stories in structure 2.

Although the story structure was not shown to influence the total reaction times, it poses an important question. In what ways does the story structure influence the question difficulty? What contextual factors of the question influence the difficulty of the questions? And is the influence on total reaction time caused by factors like the Order of Knowledge and self-reflexivity larger for more difficult story structures? In this thesis, some first steps were made to find one of these factors.

Although these first steps did not yet lead to clear answers, the results are promising enough to continue. The approach of this thesis, of categorizing the stories by their corresponding Kripke models, can be extended. Other stories with similar or even simpler models could be added.

Furthermore, factual questions could be added to the experimental setup. The difficulty of these questions would not be influenced by higher-order logical complexity, but by word choice and maybe by specific, non-structural aspects of the corresponding stories.

Therefore, the effects of the Kripke models on the question difficulty could be distinguished from other effects.

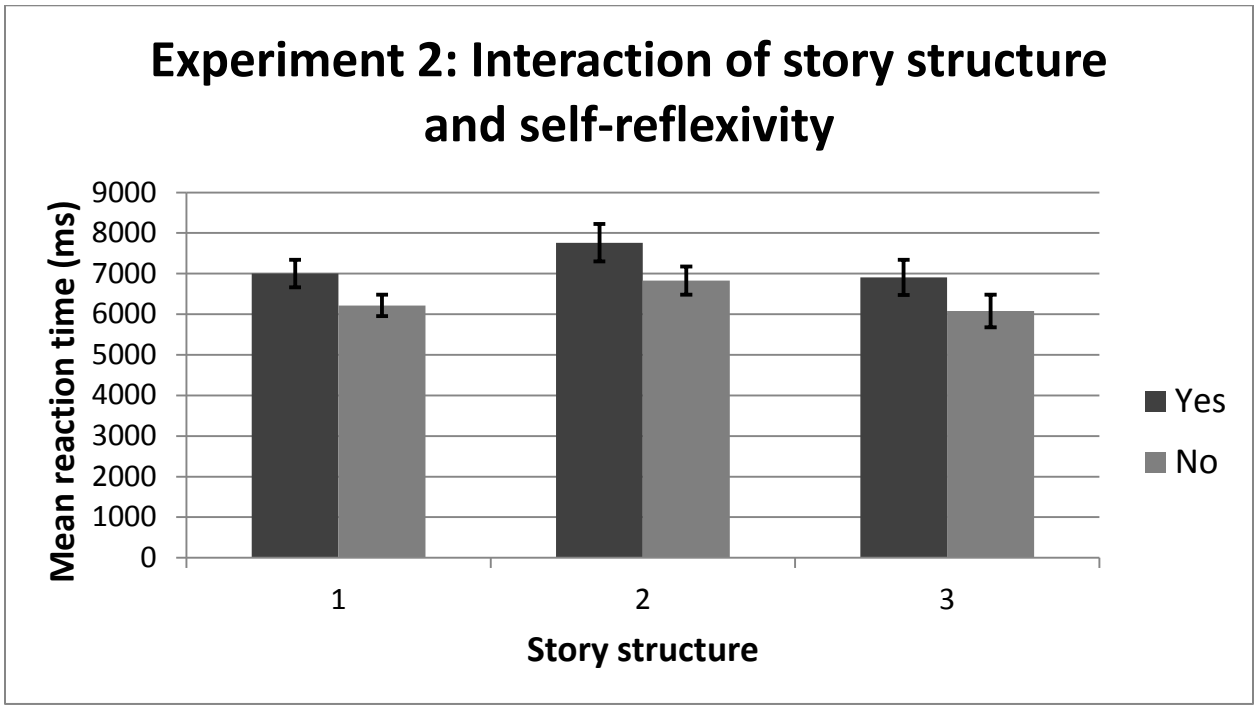


Figure 8: the influence of story structure and self-reflexivity on the total reaction times in Experiment 2

9.2 Number of fixations

In both experiments, I recorded the fixations of the participants while they were answering the questions, as measured by the eye-tracker. The idea was to look at the number of fixations during the decision times and to infer whether the participants were rereading parts of the questions.

The participants did seem to do this for several questions, but the number of fixations was so low during the decision times that the results were not significant. In addition to this, these fixations showed a similar pattern as the decision times itself, as the number of fixations per second was more or less constant for participants.

Therefore, I decided to leave out the analysis of number of fixations during the total reaction times, reading times, and decision times from this thesis. For future work, the input of the eye-tracker may be very useful in determining the border between reading and deciding. However, the number of fixations does not seem to add any information for this kind of experiments.

10 Conclusions

One important question that has been investigated thoroughly is whether certain animals (Premack & Woodruff, 1978; Penn & Povinelli, 2007) or autistic children (Baron-Cohen et al., 1985), and also at what age children develop ToM reasoning abilities (Leslie, 1987). All this research led to the insight that ToM reasoning is not an absolute ability. Even though some animals or young children show some behavior that suggests ToM reasoning, it is unclear whether they understand the mental states of others (Penn & Povinelli, 2007), or that this behavior just occurred by chance or as a result of behavior rules (Heyes, 1998).

Adult human beings have often been ignored as a research subject (Apperly, 2011), but some restrictions were discovered in their ToM reasoning (Birch & Bloom, 2007). For example, the curse of knowledge: when human adults are worse at reasoning about other people's beliefs when their truth differs from their own truth (Birch & Bloom, 2007).

There has been a lot of research on eye movements in reading and information processing (Rayner, 1998). There has also been some research about what eye movements can say about ToM (Meijering, van Rijn, Taatgen, & Verbrugge, 2012). Researching eye movements during a ToM reading task was therefore expected to be insightful. In this way, the knowledge about eye movements during reading could be used to learn more about the information processing of ToM tasks.

The classic ToM experiment can be described as stories with questions. It has often been used for children in different age groups to determine their ToM reasoning capabilities (see Section 2.1). In this thesis, the accuracy to the questions was expected to be near perfect, as it turned out to be. The measurements of the difficulty of the question were the reaction times. These reaction times were divided within three groups: the total reaction times, the reading times, and the decision times.

In this thesis, two experiments were done. In both these experiments, the experimenter asked participants to answer questions to stories. Those questions asked participants about the beliefs of the characters in the stories. The participants themselves were also characters in the story represented by 'you'.

The first experiment seemed to be too easy for adult participants and the participants also seemed to be thinking while they were still reading the questions. Therefore, the second experiment was made more difficult by involving the concurrent working memory task. This was done by increasing the work-load with a simple higher/lower-game as a dual task.

The stories were divided into three groups, based on the corresponding Kripke models. These models were the models representing the knowledge of the characters in the story at the end of the story. Although it was expected that an increased number of Kripke states would increase the decision times, this effect was not shown.

Another factor was investigated that combined the story structure and the specific question, namely; the number of states of the Kripke model that actually needed to be visited to answer the question. However, this factor did not seem to have an effect on either the total reaction times, the reading times, nor the decision times in either of the experiments.

The length of the question was one of the three properties of the questions itself to possibly influence the reaction times. The hypothesis was that longer questions would increase the reading times but not the decision times. This hypothesis was confirmed by both experiments.

Another property of the question was the Order of Knowledge (OoK). This reflects the degree of ToM necessary to answer the question. The expectation was that questions with a higher ToM would have longer decision times, but that this factor would not influence the reading times. However, in experiment 1, questions with higher OoK were found to have larger average reading times, but no influence of OoK was found on the decision times. In experiment 2, the results were even more surprising. The questions with higher OoK were found to have larger average total reaction times, but neither the reading times nor the decision times were influenced.

The last property inherent to the question was the self-reflexivity. This is a new term, invented for the purposes of this thesis. A question was considered to be self-reflexive when it asked about the representation of another character of the beliefs of the participant herself or himself. An example of a self-reflexive question is “Does Nina think that *you* think that the chocolate bar is on the table?”. A non-self-reflexive question is: “Does Nina think that *Levi* thinks that the chocolate bar is on the table?”.

For self-reflexive questions, the responder knows the truth about his/her beliefs. His/her knowledge of the truth was expected to interfere with someone else’s beliefs about this truth. Human adults need more time to assess situations where they need to take some else’s perspective (Wimmer & Perner, 1983). The self-reflexivity of the question was, therefore, expected to increase the decision times, but not the reading times. In Experiment 1, no effects of self-reflexivity were found, because the differences in decision times were too small to be distinguished. In experiment 2, the decision times were found to be solely dependent on the self-reflexivity and not to any other of the factors investigated. The questions that required self-reflexivity had significantly larger average decision times than the other questions. However, the self-reflexivity did not influence the reading times, or the total reaction times.

It was interesting that in Experiment 2, the self-reflexivity was found to increase the decision times, but not the total reaction times. On the other hand, the Order of Knowledge was found to influence the total reaction times, but neither the reading times nor the decision times. So therefore, those two factors are probably related. One explanation would be that the OoK did not cause the increase in difficulty of the question, but was merely a summation of two other effects: the self-reflexivity and the number of foreign names in the questions, resulting in slightly longer reading times. However, the increased reading times were not significantly shown by the experiments in this thesis.

11 List of abbreviations

OoK
Order of knowledge

OoG
Order of grammar

SR
Self-reflexivity

ToM
Theory of mind

REML
Relativized (or restricted, or residual or reduced) maximum likelihood

LME model
Linear mixed-effects model

12 References

Akaike, H. (1974). A new look at the statistical model identification. *Automatic Control, IEEE Transactions on, 19*(6), 716-723.

Apperly, I. (2011). *Mindreaders; the Cognitive Basis of "Theory of Mind"* (1st ed.). East Sussex, United Kingdom: Psychology Press.

Apperly, I. A., & Butterfill, S. A. (2009). Do humans have two systems to track beliefs and belief-like states? *Psychological Review, 116*(4), 953-970. doi:10.1037/a0016923

Baayen, R. H. (2008). *Analyzing linguistic data: A practical introduction to statistics*. Cambridge: Cambridge University Press.

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language, 59*(4), 390-412.

Baron-Cohen, S., Leslie, A. M., & Frith, U. (1985). Does the autistic child have a theory of mind? *Cognition, 21*(1), 37-46. doi:10.1016/0010-0277(85)90022-8

Bates, D., Maechler, M., & Dai, B. (2008). The lme4 package. *Computer Software Manual. Retrieved from http://cran.r-Project.org/web/packages/lme4/lme4. Pdf.*

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

Bowler, D. (1997). Reaction times to mental state and non-mental state questions in false belief tasks by high-functioning individuals with autism. *European Child & Adolescent Psychiatry, 6*(3), 160-165.

Fischer, B., & Weber, H. (1993). Express saccades and visual-attention. *Behavioral and Brain Sciences, 16*(3), 553-567.

Gallese, V. (2007). Before and below 'theory of mind': Embodied simulation and the neural correlates of social cognition. *Philosophical Transactions of the Royal Society B-Biological Sciences, 362*(1480), 659-669. doi:10.1098/rstb.2006.2002

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

Gazzola, V., & Keysers, C. (2009). The observation and execution of actions share motor and somatosensory voxels in all tested subjects: Single-subject analyses of unsmoothed fMRI data. *Cerebral Cortex, 19*(6), 1239-1255. doi:10.1093/cercor/bhn181

Gettier, E. L. (1963). Is justified true belief knowledge? *Analysis,* 121-123.

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

Gurka, M. (2006). Selecting the best linear mixed model under REML. *American Statistician, 60*(1), 19-26. doi:10.1198/000313006X90396

Harville, D. (1977). Maximum likelihood approaches to variance component estimation and to related problems. *Journal of the American Statistical Association, 72*(358), 320-338. doi:10.2307/2286796

Heyes, C. M. (1998). Theory of mind in nonhuman primates. *Behavioral and Brain Sciences, 21*(01), 101-114.

Hogrefe, G., Wimmer, H., & Perner, J. (1986). Ignorance versus false belief: A developmental lag in attribution of epistemic states. *Child Development,* 567-582.

Jacob, R. J. K. (1991). The use of eye-movements in human-computer interaction techniques - what you look at is what you get. *ACM Transactions on Information Systems, 9*(2), 152-169. doi:10.1145/123078.128728

Just, M., & Carpenter, P. (1980). A theory of reading - from eye fixations to comprehension. *Psychological Review, 87*(4), 329-354. doi:10.1037/0033-295X.87.4.329

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

Leslie, A. (1987). Pretense and representation - the origins of theory of mind. *Psychological Review, 94*(4), 412-426.

Lin, S., Keysar, B., & Epley, N. (2010). Reflexively mindblind: Using theory of mind to interpret behavior requires effortful attention. *Journal of Experimental Social Psychology, 46*(3), 551-556. doi:10.1016/j.jesp.2009.12.019

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability,* Berkeley: University of California Press. *, 1*(281-297) 14.

Meijering, B., van Rijn, H., Taatgen, N. A., & Verbrugge, R. (2012). What eye movements can tell about theory of mind in a strategic game. *PloS ONE, 7*(9), e45961. doi:10.1371/journal.pone.0045961

Meyer, J. -. C., & van der Hoek, W. (2004). *Epistemic Logic for AI and Computer Science* Cambridge University Press, Cambridge.

Onishi, K. H., & Baillargeon, R. (2005). Do 15-month-old infants understand false beliefs? *Science (New York, N.Y.), 308*(5719), 255-258. doi:308/5719/255

Penn, D. C., & Povinelli, D. J. (2007). On the lack of evidence that non-human animals possess anything remotely resembling a 'theory of mind'. *Philosophical Transactions of the Royal Society B-Biological Sciences, 362*(1480), 731-744. doi:10.1098/rstb.2006.2023

Premack, D., & Woodruff, G. (1978). Does the chimpanzee have a theory of mind? *Behavioral and Brain Sciences, 1*(4), 515-526.

Rayner, K. (1998). Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin, 124*(3), 372-422. doi:10.1037/0033-2909.124.3.372

Rayner, K., & Duffy, S. (1986). Lexical complexity and fixation times in reading - effects of word-frequency, verb complexity, and lexical ambiguity. *Memory & Cognition, 14*(3), 191-201. doi:10.3758/BF03197692

Samson, D., Apperly, I. A., Braithwaite, J. J., Andrews, B. J., & Scott, S. E. B. (2010). Seeing it their way: Evidence for rapid and involuntary computation of what other people see. *Journal of Experimental Psychology-Human Perception and Performance, 36*(5), 1255-1266. doi:10.1037/a0018729

Steinman, R., Haddad, G., Skavensk, A., & Wyman, D. (1973). Miniature eye-movement. *Science, 181*(4102), 810-819. doi:10.1126/science.181.4102.810

Stenning, K., & van Lambalgen, M. (2007). Logic in the study of psychiatric disorders: Executive function and rule-following. *Topoi, 26*(1), 97-114.

Stich, S., & Ravenscroft, I. (1994). What is folk psychology? *Cognition, 50*(1-3), 447-468. doi:10.1016/0010-0277(94)90040-X

Surtees, A. D., & Apperly, I. A. (2012). Egocentrism and automatic perspective taking in children and adults. *Child Development, 83*(2), 452-460.

Szymanik, J., & Zajenkowski, M. (2010). Comprehension of simple quantifiers: Empirical evaluation of a computational model. *Cognitive Science, 34*(3), 521-532.

van der Hoek, W., & Verbrugge, R. (2002). Epistemic logic: A survey. *Game Theory and Applications, 8*, 53.

van Ditmarsch, H., van Eijck, J., Sietsma, F., & Wang, Y. (2012). On the logic of lying. *Games, Actions, and Social Software*, 53, 41-72, Springer Berlin Heidelberg.

Wimmer, H., & Perner, J. (1983). Beliefs about beliefs - representation and constraining function of wrong beliefs in young childrens understanding of deception. *Cognition, 13*(1), 103-128. doi:10.1016/0010-0277(83)90004-5

Appendix A – Stories

In this appendix, the stories are printed that were used for the experiment. All stories are in Dutch.

Story 0 – practice – structure 1b

Stel je voor dat je samen met Nina en Levi een auto gaat kopen. Na lang overleggen zijn jullie het erover eens dat de blauwe auto het mooiste is. Dus je besluit die te kopen. De volgende dag keer je alleen terug naar de showroom voor een proefrit. De auto blijkt niet lekker te rijden, dus je koopt een rode auto in plaats van de blauwe. Je rijdt direct daarna langs Nina om haar de auto te laten zien. Levi weet nog van niets.

Story 1 – structure 1a

Stel je voor dat je samen met Nina en Levi in de supermarkt bent. Jullie besluiten met zijn allen om vla met aardbeien als toetje te eten. Daarom doen jullie vanillevla en aardbeien in jullie mandje. Nina wordt door haar vriendin gebeld en gaat naar haar toe. In haar afwezigheid veranderen jullie van mening en vervangen de vanillevla door chocoladevla. Op weg naar huis denken Levi en jij al aan het heerlijke toetje.

Story 2 – structure 1a

Stel je voor dat je samen met Levi in zijn huiskamer bent. Nina komt binnen met een reep chocola voor Levi, omdat hij jarig is geweest. Levi legt de reep chocola op de tafel. Levi verlaat de kamer om boodschappen te gaan doen, Nina blijft bij jou. Je besluit de reep chocola achter de kast te verstoppen. Nina ziet jou de chocola verstoppen en je ziet haar verbaasd kijken. Je bent ondertussen moe geworden, dus je gaat naar je eigen huis.

Story 3 – structure 2

Stel je voor dat je samen met Levi uitgenodigd bent voor Nina haar verjaardag. Jullie hebben haar wijs gemaakt dat jullie saaie studieboeken voor haar gekocht hebben. Stiekem hebben jullie een fiets gekocht. Tijdens een lunchafspraak komt het aankomende verjaardagscadeau ter sprake. Nina baalt nogal dat zij suffe studieboeken krijgt. Jij stelt haar gerust door te zeggen dat zij eigenlijk een fiets krijgt. Je vraagt Nina wel om niets tegen Levi te vertellen en Nina belooft zich hieraan te houden.

Story 4 – structure 2

Stel je voor dat je samen met Nina en Levi in een bar zit. Jullie besluiten om een glaasje tequila te gaan drinken. Levi gaat even naar de wc en Nina en jij vervangen de tequila door water. Levi kan natuurlijk niet zien dat het water is, maar jij fluistert hem dat stiekem in zijn oor. Nina heeft geen idee dat jij het Levi verteld hebt.

Story 5 – structure 1b

Stel je voor dat je samen met Levi en Nina een paspop moet aankleden. Na lang beraad doen jullie de paspop een rode jurk aan. Daarna gaan Levi en Nina naar huis en blijf jij achter om af te sluiten. Als zij weg zijn kleed je de paspop opnieuw aan, ditmaal in een blauw spijkerpak. Je kijkt verschrikt op als Levi opeens binnenkomt omdat hij zijn telefoon was vergeten. Hij vindt het blauwe spijkerpak gelukkig ook mooier. Samen sluiten jullie af en gaan naar huis.

Story 6 – structure 1b

Stel je voor dat je samen met Levi en Nina gebak gaat eten. Omdat jullie allemaal geen geld hebben, besluiten jullie om simpele vanillecake te nemen. Jij hebt Nina en Levi niet verteld dat je net je loon binnen hebt gekregen. Jij besluit om ze te trakteren op een heerlijke chocoladetaart. Als je terug bij jullie tafel komt, met de chocoladetaart, is Levi weg om nog even snel een studieboek te kopen. Nina kijkt superblij en verlekkerd naar de chocoladetaart en neemt direct een hapje om te proeven.

Story 7 – structure 3

Stel je voor dat je samen met Levi en Nina een potje poker speelt. Je bekijkt je kaarten en je hebt twee vijven. Dan ga je naar de wc. Als je terugkomt gaat Nina even bier halen. Ondertussen fluistert Levi je toe dat Nina en hij jou stiekem twee azen gegeven hebben. Nina komt terug en jullie gaan verder met poker. Nina heeft niets gemerkt van het onderonsje tussen jou en Levi.

Story 8 – structure 3

Stel je voor dat je samen met Nina en Levi naar de dierentuin gaat. Om vijf uur mogen jullie de tijgers voeren. Jij gaat alleen, zonder Nina en Levi, naar het insectenhuis. Daarna kom je Nina toevallig tegen. Zij vertelt je dat zij samen met Levi het voeren verzet heeft naar drie uur. Dus lopen jullie samen door. Levi is ondertussen nog nergens te bekennen.

Appendix B – Questions

In this section, all possible questions are listed for one story. For the other stories the question were constructed in the same way. For the story and the first two questions English translations are provided. With this information, the other questions should be readable to English speakers.

Story in Dutch: Stel je voor dat je samen met Levi in zijn huiskamer bent. Nina komt binnen met een reep chocola voor Levi, omdat hij jarig is geweest. Levi legt de reep chocola op de tafel. Levi verlaat de kamer om boodschappen te gaan doen, Nina blijft bij jou. Je besluit de reep chocola achter de kast te verstoppen. Nina ziet jou de chocola verstoppen en je ziet haar verbaasd kijken. Je bent ondertussen moe geworden, dus je gaat naar je eigen huis.

Story in English: Imagine that you (character A) are with Levi (character C) in his living room. Nina (character B) enters with a chocolate bar for Levi, because he had his birthday. Levi puts the chocolate bar on the table. Levi leaves the room to do groceries, Nina stays with you. You decide to hide the chocolate bar behind the closet. Nina sees that you hide the chocolate and you see her surprised look. Then you are tired and go home.

Condition B

Denkt Levi dat de chocola op tafel ligt? / Does Levi think that the chocolate is on the table?

Denkt Levi dat de chocola achter de kast ligt? / Does Levi think that the chocolate is behind the closet?

Denkt Nina dat de chocola op tafel ligt?

Denkt Nina dat de chocola achter de kast ligt?

Condition AB

Denk jij dat Levi denkt dat de chocola op tafel ligt?

Denk jij dat Levi denkt dat de chocola achter de kast ligt?

Denk jij dat Nina denkt dat de chocola op tafel ligt?

Denk jij dat Nina denkt dat de chocola achter de kast ligt?

Condition BA

Denkt Levi dat jij denkt dat de chocola op tafel ligt?

Denkt Levi dat jij denkt dat de chocola achter de kast ligt?

Denkt Nina dat jij denkt dat de chocola op tafel ligt?

Denkt Nina dat jij denkt dat de chocola achter de kast ligt?

Condition BC

Denkt Levi dat Nina denkt dat de chocola op tafel ligt?

Denkt Levi dat Nina denkt dat de chocola achter de kast ligt?

Denkt Nina dat Levi denkt dat de chocola op tafel ligt?

Denkt Nina dat Levi denkt dat de chocola achter de kast ligt?

Condition ABA

Denk jij dat Levi denkt dat jij denkt dat de chocola op tafel ligt?

Denk jij dat Levi denkt dat jij denkt dat de chocola achter de kast ligt?

Denk jij dat Nina denkt dat jij denkt dat de chocola op tafel ligt?

Denk jij dat Nina denkt dat jij denkt dat de chocola achter de kast ligt?

Condition ABC

Denk jij dat Levi denkt dat Nina denkt dat de chocola op tafel ligt?

Denk jij dat Levi denkt dat Nina denkt dat de chocola achter de kast ligt?

Denk jij dat Nina denkt dat Levi denkt dat de chocola op tafel ligt?

Denk jij dat Nina denkt dat Levi denkt dat de chocola achter de kast ligt?