University of Groningen

Accelerating the Development of Second-Order False Belief Reasoning

Arslan, Burcu; Verbrugge, Rineke ; Taatgen, Niels; Hollebrandse, Bart

Published in:
Child Development

DOI:
[10.1111/cdev.13186](https://doi.org/10.1111/cdev.13186)

IMPORTANT NOTE: You are advised to consult the publisher's version (publisher's PDF) if you wish to cite from it. Please check the document version below.

Document Version
Publisher's PDF, also known as Version of record

Publication date:
2020

[Link to publication in University of Groningen/UMCG research database](#)

Citation for published version (APA):

Arslan, B., Verbrugge, R., Taatgen, N., & Hollebrandse, B. (2020). Accelerating the Development of Second-Order False Belief Reasoning: A Training Study With Different Feedback Methods. *Child Development*, 91(1), 249-270. <https://doi.org/10.1111/cdev.13186>

Copyright

Other than for strictly personal use, it is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license (like Creative Commons).

The publication may also be distributed here under the terms of Article 25fa of the Dutch Copyright Act, indicated by the "Taverne" license. More information can be found on the University of Groningen website: <https://www.rug.nl/library/open-access/self-archiving-pure/taverne-amendment>.

Take-down policy

If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

Downloaded from the University of Groningen/UMCG research database (Pure): <http://www.rug.nl/research/portal>. For technical reasons the number of authors shown on this cover page is limited to 10 maximum.

Accelerating the Development of Second-Order False Belief Reasoning: A Training Study With Different Feedback Methods

Burcu Arslan, Rineke Verbrugge, Niels Taatgen, and Bart Hollebrandse
University of Groningen

One-hundred-six 5-year-olds' ($M_{\text{age}} = 5;6$; $SD = 0.40$) were trained with second-order false belief tasks in one of the following conditions: (a) *feedback with explanation*; (b) *feedback without explanation*; (c) *no feedback*; (d) *active control*. The results showed that there were significant improvements in children's scores from pretest to posttest in the three experimental conditions even when children's age, verbal abilities, or working memory scores were controlled for. The training effect was stable at a follow-up session 4 months after the pretest. Overall, our results suggest that 5-year-olds' failures in second-order false belief tasks are due to lack of experience and that they can be helped over the threshold by exposure to many stories involving second-order false belief reasoning, including why questions.

Children's everyday social competence is dependent on reasoning about others' mental states, such as beliefs, desires, or intentions, which can be different from their own—called theory of mind (ToM; Premack & Woodruff, 1978). A litmus test of children's ToM is called the false belief task (Wellman, 1990; Wimmer & Perner, 1983). The success on first-order false belief tasks indicates the ability to consider another agent's (false) belief, which is different from reality, and use that information to predict or interpret that agent's behavior. For example, a child could reason that Marieke (falsely) believes that the chocolate is in the drawer and predict that Marieke will look into the drawer when she wants to eat the chocolate. Many studies have shown that, before the age of 4, most children cannot pass verbal first-order false belief tasks: They predict

another agent's behavior based on their own perspective, to which we refer as zero-order ToM reasoning (Wellman, Cross, & Watson, 2001).

Interestingly, once children are able to pass first-order false belief tasks, it takes them between one and three more years to use this false belief reasoning recursively by attributing a false belief to a protagonist who is attributing a belief to another character in the story (Perner & Wimmer, 1985; Sullivan, Zaitchik, & Tager-Flusberg, 1994). For example, "Marieke (falsely) *believes that* Kevin *believes that* the chocolate is in the drawer." This level of false belief reasoning is called second-order false belief reasoning. It has been argued that while first-order false belief reasoning indicates an important cognitive advancement in terms of belief understanding, second-order false belief reasoning indicates another important advancement, which is recursive belief reasoning (Perner & Wimmer, 1985).

Whereas first-order false belief reasoning is associated with social skills, such as deception (Sodian, Taylor, Harris, & Perner, 1992) and pretend play (Leslie, 1987), second-order false belief reasoning is consequential for more advanced aspects of children's everyday social competence, such as idiom understanding (Caillies & Le Sourn-Bissaoui, 2013), irony understanding (Filippova & Astington, 2008), and reasoning about evidence (Astington, Pelletier, & Homer, 2002). As a concrete example, to successfully maintain a strategic lie, the liar has to reason

Burcu Arslan is now at Cognitive and Technology Sciences (CATS) Center, Educational Testing Service (ETS), 660 Rosedale Rd., Princeton, NJ 08541.

We are grateful to the Netherlands Organization for Scientific Research for Vici grant NWO-277-80-01, awarded to Rineke Verbrugge, and to the European Research Council for ERC-StG grant 283597, awarded to Niels Taatgen. We are also thankful to the managers and teachers of Joseph Haydn School in Groningen, to the children who participated, and to the children's families who allowed us to carry on this study. Finally, we thank Avik Kumar Maitra for the illustrations of the stories, Bea Valkeier for being the voice of the stories, Maximilian Seidler for running the pilot study and for writing the code of the experiment, and student assistants Marten Schutten, Amélie la Rio, Bram Wiggers, Marlies Hoekstra, and Merel Wiersma for testing the children. We also thank the four anonymous reviewers for their many helpful suggestions.

Correspondence concerning this article should be addressed to Burcu Arslan, Department of Artificial Intelligence, Faculty of Science and Engineering, University of Groningen, P.O. Box 407, 9700 AK Groningen, The Netherlands. Electronic mail may be sent to barslan@ets.org.

about what the listener *knows* about what the liar *knows*, requiring second-order ToM (Talwar & Lee, 2008). Although there are many studies on children's development of ToM until the age of 4, we have relatively less knowledge about children's development of second-order ToM (Miller, 2009, 2012), which is important for more advanced social skills.

Why does it take children another couple of years to pass second-order false belief tasks once they are able to attribute a false belief to another agent? In line with the first-order ToM literature, two possible explanations have been proposed, namely conceptual change and complexity (Miller, 2009, 2012). Whereas a pure conceptual change explanation of first-order false belief reasoning suggests that children need to understand that different agents may have different (false) beliefs from their own, a pure conceptual change explanation of second-order false belief reasoning suggests that children need to understand that beliefs can have other beliefs and not just events in the world as their content, meaning that they start to reason about beliefs recursively—for example, “*Marieke (falsely) believes that Kevin believes that the chocolate is in the drawer,*” not just “*Kevin (falsely) believes that the chocolate is in the drawer*”.

On the other hand, a pure complexity explanation suggests that the higher complexity of second-order false belief tasks adds further demands on working memory, as does the linguistic complexity of the stories and the questions, in comparison to first-order false belief tasks. Most studies on second-order false belief reasoning investigated the roles of executive functions and language with correlational studies. These studies indicate that executive functions and recursive language are correlated with children's development of second-order false belief reasoning (Arslan, Hohenberger, & Verbrugge, 2017; de Villiers, Hobbs, & Hollebrandse, 2014; Hasselhorn, Mähler, & Grube, 2005; Hollebrandse, Hobbs, de Villiers, & Roeper, 2008; Hollebrandse, van Hout, & Hendriks, 2014; Perner, Kain, & Barchfeld, 2002).

In addition to a possible conceptual change and necessary complexity-related developments occurring before the age of 5, we propose to add the importance of experience. We argue that most 5- to 6-year-olds, who are on the brink of passing second-order false belief tasks, understand that they can use false belief reasoning recursively (i.e., conceptual change), have sufficient language abilities to understand second-order false belief questions, and have the cognitive skills required to carry out

second-order false belief reasoning (i.e., complexity). However, they are not used to apply second-order false belief reasoning frequently in daily life; therefore, they lack experience in realizing that second-order false belief reasoning is needed.

A support for the role of experience comes from previous studies on adults' ToM reasoning. We can safely assume that adults already know that beliefs can be used recursively, and they have sufficient language and executive functions to use second-order ToM reasoning. Studies on adults' ToM reasoning in strategic games have shown that adults start applying lower levels of ToM reasoning and with accumulated evidence slowly increment their level of ToM reasoning (e.g., from first-order ToM to second-order ToM) when it is necessary (Goodman et al., 2006; Meijering, Taatgen, Van Rijn, & Verbrugge, 2014). Similar to these findings, we hypothesized that when 5-year-olds are trained with many second-order false belief tasks by providing explicit feedback (correct/wrong) with or without explanation, they can revise their first-order false belief reasoning to correct second-order false belief reasoning (see also Goodman et al., 2006; Arslan, Taatgen, & Verbrugge, 2017 for computational cognitive models on the role of experience in children's development of first-order and second-order false belief reasoning, respectively).

In the following subsection, we first review some of the previous training studies of ToM. Subsequently, we explain the details of our training study and the specifics of our hypothesis.

Training Studies on ToM: Related Work

Several training studies have shown that it is possible to accelerate preschool children's development of first-order ToM with a moderately strong effect size with a relatively short training program (see Kloo & Perner, 2008 for a review, and Hofmann et al., 2016 for a meta-analysis). These studies have shown the importance of explicit feedback (Clements, Rustin, & McCallum, 2000; Melot & Angeard, 2003; Slaughter, 1998; Slaughter & Gopnik, 1996), explicit feedback with self-generated explanations (Amsterlaw & Wellman, 2006; Guajardo, Petersen, & Marshall, 2013), conversations about mental states (Appleton & Reddy, 1996; Hale & Tager-Flusberg, 2003; Lohmann & Tomasello, 2003; Ornaghi, Brockmeier, & Gavazzi, 2011), as well as cognitive control (Kloo & Perner, 2003) and language (Hale & Tager-Flusberg, 2003; Lohmann & Tomasello, 2003) on the development of first-order false belief reasoning.

Given that children's ToM development goes beyond first-order false belief reasoning and continues to develop after they reach the age of 4, a couple of previous training studies focused on children's ToM development beyond the preschool years (Bianco, Lecce, & Banerjee, 2016; Lecce, Bianco, Devine, & Hughes, 2014). In these studies, 9- to 10-year-old children were trained with a conversation-based approach in order to investigate the efficacy of conversations about mental states in children's development of more advanced ToM reasoning. Because most children around the age of 9 already pass second-order false belief tasks, children were trained with a more advanced and naturalistic ToM task—a version of the Strange Stories task (Happé, 1994)—in which their ability to make inferences about mental states in nonliteral statements was assessed (e.g., double bluffs, white lies, misunderstandings). During the training sessions, children participated in a group conversation about the strange stories and got corrective feedback and further explanations. In the control condition, children had to reason about similar stories, however, involving physical events instead of mental state reasoning. The findings showed that children's performance from pretest to posttest significantly improved for children in the experimental condition compared to the children in the control condition, and this improvement was stable over 2 months.

In summary, these training studies on more advanced ToM with older children support the previous findings from the first-order false belief reasoning literature showing that it is possible to accelerate children's ToM development with explicit feedback and further explanations. However, in these studies children were not trained in a condition in which they would perform advanced ToM tasks with only the feedback of "Correct" or "Wrong" together with the correct answer, or without any feedback at all. Therefore, it is still unknown whether children would still improve in those conditions. Moreover, as far as we know, there is no literature on the role of different types of feedback on second-order false belief tasks for children between the ages 4 and 9, especially those children on the brink of developing second-order false belief reasoning.

The Current Study

Our training study fills the above-mentioned gaps in the literature by training 5- to 6-year-old children, who are on the brink of passing second-order false belief tasks, with second-order false belief tasks by providing different types of feedback

in the following conditions: (a) *feedback with explanation*: by providing feedback "Correct" or "Wrong" together with the correct answer and further explanations about the reason why it is the answer; (b) *feedback without explanation*: by providing feedback "Correct" or "Wrong" together with the correct answer but without further explanations; (c) *no feedback* and an (d) *active control* condition in which children were trained with neutral stories that do not involve any level of false belief reasoning.

Moreover, few of the previous training studies tested children again in a follow-up session a couple of months after the posttest session in order to assess whether children's improvements were stable over time (Hofmann et al., 2016). Importantly, the methodology of our training study covers and extends the important suggestions of Hofmann et al. (2016) meta-analysis of the training studies of ToM, by: (a) controlling for working memory, verbal abilities, and age; (b) testing children in a follow-up session 4 months after the pretest session; (c) using an active control condition; and (d) controlling for children's pattern learning instead of reasoning about others' minds by testing them with second-order true belief stories.

In addition to second-order false belief tasks, we tested children with a working memory task in order to control for a possible effect of the training on working memory (see Arslan, Hohenberger, et al., 2017; Perner et al., 2002 for the role of working memory in children's development of second-order ToM; but see also Hasselhorn et al., 2005 for no significant correlation between the working memory span score and children's second-order false belief score when verbal abilities and age were controlled for).

We have two specific hypotheses about the possible effect of training children using feedback with explanation and feedback without further explanations. First, based on the previous first-order and advanced ToM training studies, we expect that children who are in the feedback with explanation condition will show an improvement in their second-order false belief scores from pretest to posttest session and that their improvement will be greater than those of the children in the other conditions. Second, we expect that children who are in the feedback without explanation condition will also show improvement from pretest to posttest sessions and that their improvement will be greater than that of the children in the active control condition.

The rationale for these hypotheses is based on our assumption that 5-year-olds are able to attribute second-order false belief reasoning, however, they

lack experience in using it when they are asked second-order false belief questions (e.g., “Where does Marieke think that Kevin will look for the chocolate?”). Therefore, the feedback without explanation together with the correct answer provides evidence for children to realize that second-order false belief reasoning is needed. However, second-order false belief reasoning would be more likely when feedback with explanations is provided than feedback without explanations, because the explanations provide further evidence for the correct level of reasoning (Arslan, Taatgen, et al., 2017). We provide a more detailed discussion about our hypotheses in the light of our findings in the Discussion section.

Method

Participants

One hundred nineteen 5- to 6-year-old ($M_{\text{age}} = 5;6$; $SD = 0.40$) children were recruited from a primary school with predominantly upper-middle-class Dutch families living in a university town with plenty of children’s activities and facilities (parks, bicycle trails etc.), namely Groningen, the Netherlands (see Table 1 for the detailed presentation of age, sex in each condition). Children’s socioeconomic status information was verbally obtained from the school management and the teachers of the individual classes in which the children were recruited. All children had Dutch as their first language, and they were students of three different teachers. To each of the four conditions we assigned equal numbers of children from the different teachers’ classes. Data collection was started in February 2014 and ended in March 2016, including the follow-up sessions, which took place 4 months after the posttest session. We sent a written parental consent form to the parents via the teachers. The children whose parents did not object to participation in the experiment and who did not have cognitive or learning difficulties were initially included. None of those children objected to participate our

training program, except for one child who started to cry during the pretest. This child was brought back to class and was not asked to participate any further.

Children were pretested to ensure that they had not yet fully developed second-order false belief reasoning. Thirteen children were excluded from the study, as follows: Nine of these 13 children (aged 5;0, 5;3, 5;3, 5;4, 5;5, 5;8, 5;8, 5;8, 6;1) were already good at second-order false belief reasoning and gave correct answers for all the three second-order false belief questions. Two of the 13 children (aged 5;4, 5;8) left the study before it was completed; moreover, one child was excluded due to technical problems during the experiment (aged 5;5), and one child (aged 5;1) was excluded because she was not able to answer any of the first-order false belief questions at the pretest. Thus, the analysis included the results of 106 children in three experimental conditions and one control condition.

Design

Children were randomly assigned to three different experimental conditions and one control condition: (a) feedback with explanation; (b) feedback without explanation; (3) no feedback; (4) active control. Each child was tested in five separate sessions, namely pretest, training Day 1, training Day 2, posttest, and follow up. There was at least 1 day of intermission between each pair of subsequent sessions (the pretest, training Day 1, training Day 2, and posttest sessions), and there was at least 1 week and at most 9 days of intermission between the pretest and the posttest sessions. The follow-up session was conducted 4 months after the posttest session. Each session took between 30 and 45 min. Figure 1 shows the design of the experiment.

Procedure

Children were tested individually in their school in a separate room by one of seven experimenters.

Table 1

Number of Participants, Age Range, Mean Age (Standard Deviation), Mean Verbal Ability Scores (Standard Deviation), and Mean Working Memory Score (Standard Deviation) at Pretest in Each Condition

Condition	N	Age range	M_{age} (SD)	Verbal ability (SD)	WM _{Pretest} (SD)
Feedback with explanation	23 (15 female)	5;1–6;2	5;8 (0.29)	51.00 (5.08)	2.45 (0.98)
Feedback without explanation	23 (10 female)	5;2–6;8	5;8 (0.44)	53.71 (3.62)	1.98 (1.16)
No feedback	26 (11 female)	5;2–6;8	5;4 (0.25)	53.00 (4.31)	2.25 (1.13)
Control	34 (19 female)	4;8–6;5	5;3 (0.35)	51.51 (7.18)	2.17 (1.20)

Figure 1. The design of our training study. [Color figure can be viewed at wileyonlinelibrary.com]

Because children had not yet learned how to read, all the stories and the questions were presented via the computer's speakers. All the drawings and the audio files were implemented in Psychopy2 v.1.78.01 and were presented to the children on a 15-in. Apple Inc. Cupertino, California, U.S. MacBook Pro OS X 10.10.5. In each second-order false belief story and each second-order true belief story of a certain type, we fixed the general story structure, but we changed the protagonists' gender, appearance, and name, as well as objects, locations, goals, and further context of the stories.

All experimenters were trained before running the experiment in order to follow the same instructions. After introducing themselves, experimenters told children that they were going to hear stories and answer questions that were presented via a computer. Also, they told children in the pretest, posttest, and follow-up sessions that they were going to play a counting game (i.e., the counting span task) via the computer. A child was almost always tested by the same experimenter at the pretest, training Day 1, training Day 2, posttest and follow-up sessions. As an exception, two of the experimenters were not able to attend the follow-up test for 17 children, which was 4 months after the posttest. Thus, two of the remaining five experimenters tested these 17 of 106 children for the follow-up session. Each session took approximately 30 min. All of the sessions were recorded with QuickTime's screen recording together with audio recording. After each session, children received three stickers for "doing so well."

The stories were drawn randomly without repetitions from a pool that contained 31 different second-order false belief stories, a pool of four different second-order true belief stories, and (for the control condition) a pool of 14 different neutral

stories. Drawings illustrating the story episodes were presented one by one, together with the corresponding audio recordings via computer. The drawings remained visible when children were asked the pre-recorded questions. While children were listening to the stories, experimenters remained silent; however, they pointed at the related story drawing on the screen in order to make sure that children paid attention to the stories.

As has been usual in previous studies, several control questions were asked in the course of each story in order to test that children did not have major memory and linguistic problems about the stories and the structure of the questions (Sullivan et al., 1994; Wimmer & Perner, 1983). Also, a first-order false belief question was asked at an appropriate moment in the story, before the second-order false belief question, in order to make sure that the children did not have any major problems with first-order false belief reasoning. Children gave verbal answers to the questions, directed to the experimenter. If a child gave a wrong answer for a control question or a first-order false belief question in the second-order false belief tasks or for the questions in the neutral stories that were used in the control condition, the experimenters pressed a key on the keyboard to repeat the relevant part of the story and the prerecorded questions, up to three times altogether. Also, if a child said "I don't know" for the second-order false belief question, the experimenter repeated the prerecorded second-order false belief question up to two more times. The justification question for the second-order false belief questions was asked by the audio, for example, "Why does mom say that?" If a child did not answer at first, the experimenter would repeat the question once but would not probe any further after that.

Pretest, Posttest, and Follow-Up Testing Sessions

Children were tested with a counting span task and 3 second-order false belief stories (1 “Three goals” 1 “decoy gift,” and 1 “three locations”) in a random order via computer (see Materials section for the details of the stories). The presentation of the order of the tasks and the order of the story types were randomized. Children did not get any feedback in the pretest, posttest, and follow-up sessions.

In order to test whether children can generalize what they learned in the training sessions to another type of second-order false belief task, children were not trained with “three locations” stories at the two training sessions. Moreover, in order to make our training program more robust, we trained children with two different types of second-order false belief stories at the training sessions instead of training children with only one type of second-order false belief story. All the three different types of stories test children’s second-order false belief understanding. However, although “three locations” and “three goals” stories have three possible answers, “decoy gift” stories have only two possible answers (see Materials section for examples of each story type).

Training Sessions

In the second and the third sessions (training Day 1, training Day 2), children in the three experimental conditions were trained using six different second-order false belief stories (3 “decoy gift,” and 3 “three goals”) per training session. At each training session, children were also tested with two second-order true belief stories, where each true belief story was presented after three second-order false belief stories (see subsection *Second-Order True Belief Stories* in the Materials section for the rationale for including second-order true belief stories in our training study).

In the feedback with explanation condition, the feedback “Correct” or “Wrong” together with an explanation was provided in an interactive fashion by an experimenter. For example, the explanation followed the following script for the prototype of the “three goals” stories that was explained in the Materials section (Figure 3):

Robert told his dad that he wanted to go to the zoo. Then, mom told Robert that the zoo is not open today and they can go to the swimming pool but dad did not hear that, right? That is

why dad says to grandmother that Robert thinks they are going to the zoo, right?

We trained the experimenters to give the exact same feedback for each child, and we provided them a script. Because the feedback was provided in an interactive fashion, there were small variations in the form of the feedback between and within the participants. However, very similar information was given to each child even in those cases.

In the feedback without explanation condition, only the feedback “Correct” or “Wrong” was provided, together with the correct answer (such as “a basketball”) without any further explanation. In the no feedback condition and in the control condition, children did not get any feedback.

In the control condition, in both training sessions, children were tested with seven neutral stories and questions via computer that did not involve any level of false belief reasoning. Each neutral story had approximately the same length as the second-order false belief stories and the second-order true belief stories.

Materials

Second-Order False Belief Stories

We constructed 31 different second-order false belief stories of three different types: (a) 3 “three locations” stories (i.e., 1 pretest, 1 posttest, 1 follow-up), (b) 14 “three goals” stories (from these, for each child, 3 were used on training Days 1 and 3 on training Day 2), (c) 14 “decoy-gift” stories (again from these, for each child, 3 were used on training Days 1 and 3 on training Day 2). Note that we initially aimed to train each child with 7 “three goals” stories and 7 “decoy gift” stories per training day (i.e., in total 14 different stories of each story type). However, in our pilot study prior to our training program, we realized that children started to get distracted after 45 min. Therefore, we decided to reduce the number of stories in our training program.

For all stories, children were asked a question that required second-order false belief attribution, as well as some control questions via computer and gave verbal responses. In the literature, second-order false belief questions often have two possible answers, for example, two locations. For the purpose of another study in which we focused on children’s level of ToM reasoning (i.e., zero-order, first-order) when they fail in second-order false belief

tasks at the pretest sessions, we constructed the “three locations” and “three goals” stories in such a way that our second-order false belief questions have three different possible answers, according to which we can distinguish children’s level of reasoning.

Figure 2 shows the prototype example of “three locations” stories, namely the “chocolate bar” story. Stories of this type have in common that an object is really in location A, but protagonist one falsely believes that it is still in location B; in the meantime, protagonist two falsely believes that

protagonist one falsely believes that the object is in location C. These “three locations” stories were constructed based on Flobbe, Verbrugge, Hendriks, and Krämer’s (2008) version of Hale and Tager-Flusberg’s (2003) “chocolate bar” story. There are three possible answers to be reported to the second-order false belief question “Where does Marieke think that Kevin will look for the chocolate?”, namely: (a) the correct second-order false belief answer: “the drawer,” because Marieke *thinks that* Kevin *thinks that* the chocolate is in the drawer; (b) the incorrect first-order false belief answer: “the toy

a) Kevin and Marieke are brother and sister. They are in the living room.

b) Their mother bought a chocolate bar and gives it to Kevin. Marieke doesn't get any chocolate, because she has been naughty.

c) Kevin eats some of his chocolate and puts the remainder into the drawer. He doesn't give any chocolate to Marieke. Marieke is upset that she does not get any chocolate.

d) After that, Kevin goes to help his mother in the kitchen. Marieke is alone in the room. Because she is upset, she takes the chocolate from the drawer and puts it into the toy box. While she is putting the chocolate into the toy box, Kevin is passing by the window. He sees how Marieke takes the chocolate out of the drawer and puts it into the toy box. Marieke does not see Kevin.

At this point, the prerecorded control questions “Does Kevin know that Marieke put the chocolate into the toy box?” (yes), and “Does Marieke know that Kevin saw her put the chocolate into the toy box?” (no) were asked.

e) After that, Kevin goes back to the kitchen and Marieke goes to the kitchen, as well. While Kevin and Marieke are in the kitchen, their mother goes to the living room to watch TV. While she is searching for the remote control, she sees the chocolate in the toy box. The mother is surprised that the chocolate is in the toy box. She takes the chocolate from the toy box and puts it into the TV stand. She watches TV for a while and goes to her room.

At this point of the story, the reality control question “Where is the chocolate now?” (in the TV stand) was asked.

f) Now, Kevin and Marieke go back to the living room. Kevin wants to eat some of his chocolate. He says: ‘Hmm, I would like to some chocolate’.

At this point the first-order false belief question “Where will Kevin look for the chocolate?” (in the toybox) and the justification question “Why does he look there?” were asked.

Subsequently, the second-order false belief question: “Where does Marieke think that Kevin will look for the chocolate?” (in the drawer) was asked together with the justification question “Why does she think that?”

Figure 2. The prototype “three locations” story, namely the “chocolate bar” story (Illustration ©Avik Kumar Maitra). Correct answers for the questions are provided between parentheses after each question. [Color figure can be viewed at wileyonlinelibrary.com]

box,” because Kevin actually *thinks that* the chocolate is in the toy box (this would be the correct answer to the embedded first-order question “Where will Kevin look for the chocolate?”); (c) the incorrect zero-order ToM answer: “the TV stand,” because the chocolate is in fact in the TV stand.

Figure 3 shows a prototype example of “decoy gift” stories, namely the “birthday puppy” story. Stories of this type have in common that protagonist one will really receive a gift A and protagonist one has in fact discovered the gift and correctly believes it will be A; in the meantime, however, protagonist two falsely believes that protagonist one still falsely believes that the gift will be B. “Decoy gift” stories were constructed based on

Sullivan et al.’s (1994) birthday puppy story. Unlike the “three locations” and “three goals” stories, in this story, there are two answers that the participants might report: (a) correct second-order false belief answer: “a basketball,” because mother *thinks that* Rick *thinks that* she bought a basketball; (b) incorrect zero-order ToM and first-order false belief answer: “a puppy,” because it is the real present and because Rick *thinks that* his mother bought a puppy.

Figure 4 shows a prototype example of “three goals” stories, namely the “a day out” story. Stories of this type have in common that protagonist two really has goal A, whereas protagonist one falsely believes that the goal is B; in the meantime,

a) Tonight, it's Rick's birthday and his mum wants to surprise him with a puppy. She has hidden the puppy in the basement.

b) Rick says, “Mum, I really hope you got me a puppy for my birthday”.

c) Because Rick's mother wants to surprise him with a puppy, instead of telling Rick she got him a puppy, she says “Sorry Rick, I didn't get you a puppy for your birthday. I got you a really nice basketball instead.”

At this point, the control question “What did the mother really get Rick for his birthday” (a puppy) and the first-order false belief question “What does Rick think that his mom bought for him?” (a basketball), together with the justification question “Why does Rick think that?” were asked.

d) Now, Rick says to his mother: “I am going outside to play.” On his way outside, Rick goes down to the basement to fetch his skates. In the basement, Rick finds his birthday puppy. Rick says to himself: “Wow, mum didn't get me a basketball; she really got me a puppy for my birthday.” His mother does not see that Rick goes down to the basement and finds the birthday puppy.

At this point the control question “Does Rick know that his mother got him a puppy for his birthday?” (yes) was asked.

e) Now the telephone rings, ding-a-ling! Rick's grandmother calls to find out what time the birthday party is. The mother tells grandma on the phone that she got Rick a puppy for his birthday, but that Rick doesn't know this. Then, grandma asks mum on the phone, “What does Rick think you got him for his birthday?”

Subsequently, the second-order false belief question “What does the mother say to grandma?” (a basketball), together with the justification question “Why does mum say that?” were asked.

Figure 3. The prototype example of “decoy gift” stories, namely the “birthday puppy” story (Illustration © Avik Kumar Maitra). Correct answers for the questions are provided between parentheses after each question. [Color figure can be viewed at wileyonlinelibrary.com]

Figure 4. The drawings of the prototype of "three goals" stories, namely "a day out" story (Illustration © Avik Kumar Maitra). Correct answers for the questions are provided between parentheses after each question. [Color figure can be viewed at wileyonlinelibrary.com]

protagonist two falsely believes that protagonist one falsely believes that the goal is C. "Three goals" stories include and extend the stories used in Hollebrandse et al.' (2014) study. Just like in the "three locations" stories, there are three possible answers to the second-order false belief question: (a) correct second-order false belief answer: "the zoo," because Dad *thinks that* Robert *thinks that* they will go to the zoo; (b) incorrect first-order false belief answer: "the swimming pool," because Robert *thinks that* they will go to the swimming pool; (c) incorrect

zero-order ToM answer: "the cinema," which is the real place to which they will go.

For each story, a judgment score of 1 was given for a correct answer to a second-order false belief question, and a score of 0 was given for a wrong answer. Similarly, if a child's justification answer included the correct information that one character does or does not know about the other character's history of exposure to relevant information, it was coded as correct (1 points). Otherwise, the justification was coded as incorrect (0 points). See

Appendix B for the details of scoring justification answers.

Second-Order True Belief Stories

In addition to the second-order false belief stories in training sessions, children were tested with second-order true belief stories in order to capture whether a child's response was a result of pattern learning instead of reasoning about the second-order false belief questions. "Decoy gift" stories were presented with five pictures in which the correct second-order false belief answers were always in the third picture (Figure 3), and "three goals" stories were presented with four pictures in which the correct second-order false belief answers were always in the first picture (Figure 4). Because children were trained with 6 "decoy gift" and 6 "three goals" stories during the two training sessions, children who were in the feedback with explanation or feedback without explanation conditions might have figured out this pattern and instead of using second-order false belief reasoning, they might have just reported the object in the third picture in "decoy gift" stories, and just report the object or event in the first picture in "three goals" stories as answers for second-order false belief questions, instead of attributing second-order beliefs.

Second-order true belief stories have exactly the same structure as the second-order false belief stories and a judgment question was followed by a justification question. However, the protagonist whose belief the child has to report entertains a true belief instead of a false belief. For instance, in the true belief story corresponding to the "decoy gift" story given above, the son finds his real birthday present, but the mother is also in the room and they jointly attend the present. Therefore, this time the correct answer (a puppy) to the second-order true belief question is not the same as the correct answer (a basketball) to the second-order false belief question in the corresponding false belief story, because now the mother knows that the son knows that she bought a puppy for him. Therefore, by including second-order true belief stories in our training study, we were able to capture children who had figured out the pattern for the correct answers in second-order false belief stories and did not pay attention to the story and gave the correct answers for second-order false belief stories (e.g., a basketball) as answers for second-order true belief stories.

For each story, a judgment score of 1 was given for a correct answer to a second-order true belief

question, and a score of 0 was given for a wrong answer.

Neutral Stories

Neutral stories were presented to participants in the active control condition in two training sessions (i.e., seven stories in each training day). Fourteen neutral stories that have a similar length as the second-order false belief stories and that do not involve ToM reasoning were selected from the children's book "Jip en Janneke" by Schmidt and Westendorp's (2011) illustrations. Each story was divided into two episodes and presented on the computer with two drawings from the book illustrating the episodes. After each episode, two neutral questions not involving any mental state expressions were asked about the episode of the story, in order to check if the children paid attention (see Appendix C for an example of the neutral stories and corresponding questions used in the active control condition).

Working Memory Task

As a working memory task, we chose to use a task that involves minimal language. For this reason, we used a computerized version of the counting span task (Towse, Hitch, & Hutton, 1998). In this task, cards that have red triangles and blue squares were shown on the computer screen one by one. Children were instructed to count aloud the blue squares by pointing at them and to remember their total number on each card. The experimenter told them that after they counted the targets on the first card, the next card would be shown on the screen and they should repeat the same procedure, remembering the numbers of blue squares on both cards. After being sure that children understood the instructions and practiced one two-cards trial, which was shown on paper with the help of the experimenter, the real experiment was shown on the computer.

In the first level, after two cards, the children were asked to report the total numbers of blue squares per card in the same order that the cards had been presented. Each level had three trials. If a child reported all numbers back correctly for a trial, positive feedback was provided in the form of an audio file saying "Well done!" together with a green happy smiley on the screen. If a child was not able to report all the target numbers correctly, a neutral face together with an audio "Let's try another one!" was presented. If a child correctly

reported two out of three trials at a given level, then the difficulty was increased to a higher level, meaning that the number of cards per trial was increased by one. For the scoring, we adopted the criteria of Towse et al.'s (1998) study.

Verbal Abilities

Children's verbal ability scores were taken from their school's database and used as control variables in our statistical analyses to assess children's improvements after the training sessions. These scores are part of the monitoring system, called Cito, for schools in the Netherlands. Starting from Grade 1 to Grade 8 (4- to 12-year-old), most of the children in the Netherlands are tested with the same instruments in order to assess children's progress systematically. Children's verbal abilities were tested in terms of vocabulary and answering a question after listening to a small story.

Results

In this section, we first present preliminary analyses in order to establish the equivalence of the experimental and control conditions before the training sessions. Subsequently, we present the results of the training effects from pretest to posttest. Finally, we present the generalizability and stability of the training effects by focusing on the improvements from pretest to follow-up sessions (see Appendix A for the results of nonexperimental questions).

Preliminary Analyses

Preliminary separate Kruskal-Wallis tests were run in order to establish the equivalence of the experimental and control conditions before the training sessions. No significant differences were found between the conditions in the verbal ability scores, $\chi^2(3) = 3.04$, $p = .39$, and in the working memory scores, $\chi^2(3) = 1.55$, $p = .67$. However, we found a significant difference between the conditions in children's age, $\chi^2(3) = 32.99$, $p < .001$. Further analyses have shown that there was a significant difference of age (see Table 1) between feedback with explanation and no feedback conditions ($W = 100.5$, $p < .001$) and between feedback without explanation and no feedback conditions ($W = 139$, $p < .002$), as well as between feedback with explanation and control conditions ($W = 112$, $p < .001$) and between feedback without explanation and control conditions ($W = 145.5$, $p < .001$).

Given these results, we controlled for age in all of our subsequent analyses.

Moreover, it turned out that there were no significant differences in scores between the children who were trained with different experimenters and who were in the different classes with different teachers and adding experimenters and teachers as random effects did not improve the linear mixed effect models. Therefore, we merged the data across experimenters and teachers for the rest of the analysis.

Training Effects From Pretest to Posttest Sessions

Figure 5 shows (a) the proportion of correct answers to the second-order false belief questions at pretest, posttest, and follow-up sessions, and (b) the difference in the proportions of correct answers between pretest and posttest sessions for each condition. There is a considerable improvement of children's scores from pretest to posttest in the three experimental conditions: from 31% to 68% correct in the feedback with explanation condition; from 25% to 49% correct in the feedback without explanation condition; and from 33% to 55% correct in the no feedback condition, in contrast to a small improvement in the control condition (from 29% to 35%). Moreover, children who were in the experimental conditions performed better than the children who were in the control condition in the follow-up session, which was 4 months after the pretest session. Namely, 73% answers were correct in the feedback with explanation condition; 56% in the feedback without explanation condition; 68% in the no feedback condition; compared to 46% in the control condition.

Following a similar pattern as Figure 5, Table 2 shows the percentages (and numbers) of children showing an improvement, stability, or deterioration from pretest to posttest in their (a) answers to second-order false belief questions and (b) answers to justification questions. Children's justification answers were scored by two coders independently. Interrater reliability between the coders was 89% (Cohen's $\kappa = .78$; 95% CI [.70, .86]; $p < .0005$). All discrepancies were resolved by discussion, leading to a final resolution accepted by both coders.

As can be seen from Table 2, in the feedback with explanation condition, the percentage of children who showed an improvement in their answers to second-order false belief judgment and justification questions is higher compared to the other conditions. In the feedback without explanation and no feedback conditions, children showed similar

Figure 5. (a) Proportion of correct answers to the second-order false belief questions at pretest, posttest, and follow-up sessions. (b) The difference in the proportions between pretest and posttest sessions for each condition (with experimental conditions compared to control condition). $*p = .03$, $**p = .007$ (see Table 3). [Color figure can be viewed at wileyonlinelibrary.com]

Table 2

The Percentage (and Number) of Children Showing an Improvement, Stability, or Deterioration in (a) Answers to Second-Order False Belief Question and (b) Answers to Justification Questions From Pretest to Posttest Sessions

	Improvement		Stability		Deterioration	
	(a)	(b)	(a)	(b)	(a)	(b)
Feedback with explanation (N = 23)	74 (17)	65 (15)	9 (2)	35 (8)	17 (4)	0
Feedback without explanation (N = 23)	57 (13)	52 (12)	35 (8)	39 (9)	8 (2)	9 (2)
No feedback (N = 26)	58 (15)	54 (14)	30 (8)	27 (7)	12 (3)	19 (5)
Control (N = 34)	32 (11)	29 (10)	50 (17)	62 (21)	18 (6)	9 (3)

patterns of improvements. Moreover, as we expected, children who were in one of the three experimental conditions improved much more often, rather than staying stable or deteriorating, as compared to children in the control condition.

A binominal mixed effects model was fitted on the scores with the following effects: the main

effects of and interaction between session (pretest and posttest) and condition (feedback with explanation, feedback without explanation, no feedback, control) to test for differential learning effects of the different training regiments; a three-way interaction between condition, "three locations" items and session to test whether learning on new types of items was different from old types of items; the centered age of the child, the centered scores for verbal ability, and the centered scores for working memory capacity. As random effects, we had random slopes for session per subject correlated with the random intercepts.

Table 3 presents the estimates of the coefficients (reported in log odds) and z-statistics of the model. Note that "three location" stories were not used in the training sessions. The pretest session, "three goals" stories, and control condition were used as base levels in the model (reference categories). Therefore, the rows of Table 3 should be interpreted considering these reference categories. For example, Table 3, Row 2 shows that children's scores did not significantly improve from pretest to posttest in the control condition.

Children's second-order false belief scores significantly improved in the feedback with explanation condition and in the feedback without explanation condition in contrast to children's improvements in the control condition (Rows 11 and 12). There was

Table 3
The Estimates and z-Values of the Binomial Mixed-Effects Model From Pretest to Posttests Sessions^a

		<i>B</i>	<i>SE</i>	<i>z</i>	<i>p</i>
1	(Intercept)	-1.32	.31	-4.24	< .001
2	Posttest	0.08	.43	0.18	.86
3	Feedback without explanation	-0.67	.46	-1.46	.14
4	Feedback with explanation	-0.35	.42	-0.83	.41
5	No feedback	0.18	.37	0.48	.63
6	"Decoy gift"	1.63	.25	6.51	< .001
7	"Three locations"	-0.22	.35	-0.64	.52
8	Age	0.94	.39	2.43	.02
9	Verbal ability	-0.01	.02	-0.60	.55
10	Working memory	0.08	.11	0.72	.47
11	Posttest × Feedback Without Explanation	1.51	.70	2.17	.03
12	Posttest × Feedback With Explanation	1.88	.70	2.71	.007
13	Posttest × No Feedback	1.07	.64	1.68	.09
14	Control × "Three Locations" × Posttest	0.34	.63	0.54	.59
15	Feedback Without Explanation × "Three Locations" × Posttest	-0.86	.78	-1.11	.27
16	Feedback With Explanation × "Three Locations" × Posttest	0.28	.72	0.39	.70
17	No Feedback × "Three Locations" × Posttest	-0.06	.66	-0.09	.92

^aThe pretest session, "three goals" stories, and control condition were used as base levels in the model (reference categories).

a marginally significant improvement from pretest to posttest sessions in the no feedback condition compared to the control condition (Row 13). We did not find a significant effect of children's verbal abilities (Row 9) and working memory score at pretest (Row 10) on children's second-order false belief score. Finally, as expected, there was also a significant effect of age (Row 8). Note that there are significant differences in terms of age between the feedback with explanation and control conditions as well as between feedback without explanation and control conditions. Therefore, although we incorporated age into our statistical analyses, one might still argue that the lack of significant improvements in the control condition could be due to the fact that children's age difference. In order to investigate this possibility, we compared the younger children's (age range = 4;8–5;6) improvements from pretest to posttest sessions in the experimental conditions ($N = 35$, $M_{\text{age}} = 5;3$, $SD = 0.2$) with the younger children in the active control condition ($N = 27$, $M_{\text{age}} = 5;3$, $SD = 0.2$).

The results showed that younger children who were in the experimental conditions improved more than younger children who were in the control condition (see Appendix D).

Generalizability of the Training Effect

In order to investigate the generalizability of the training effect, we focus on children's improvements from pretest to posttest sessions in "three locations" stories. Note that, unlike "three goals" and "decoy gift" stories, we did not train children with "three locations" stories during the training sessions. As can be seen in Table 3 (Rows 6 and 7), although there was no significant difference between "three goals" and "three locations" stories (Row 7), children's scores in "decoy gift" stories were significantly better than children's scores in "three goals" stories (Row 6). Note that although the chance level of correct answers for the "decoy gift" type of stories is 50%, the chance level for the "three goals" stories is 33%. Moreover, although children's improvement in "Three locations" stories in the feedback without explanation condition was not as great as the improvement in the other conditions, there were no significant differences between children's scores in the trained types of stories and children's "three locations" posttest scores in all of the conditions.

In Figure 6, we merged the stories that we used at the training sessions ("other stories"), namely "three goals" and "decoy gift" stories and compared them with "three locations" stories. As can be seen from Figure 6, for the experimental conditions, both the "three locations" stories and the "other stories" have a similar amount of increase in the proportion of correct second-order false belief answers from pretest to posttest (a rise of 29% points in "Three locations," and a rise of 27% points in "other stories" in the experimental conditions; compared to a rise of 8% points in "three locations" and a rise of 4% points in "other stories" in the control condition). The results that we presented in Table 3, together with Figure 6 show that children were able to generalize what they learned during the training sessions to another story type with which they did not train, namely "three locations" stories.

Stability of the Training Effects: Improvements From Pretest to Follow-Up Sessions

As can be seen from Figure 5, for all the conditions, children's scores on second-order false belief

Figure 6. The comparison of children's improvements in "three locations" versus "other stories" story types of second-order false belief stories from pretest to posttest sessions. The orange dotted horizontal line represents the chance level (average of 33% and 50%) for the "decoy gift" and "three goals" stories, while the purple dotted line represents the chance level (33%) for the "three locations" stories. [Color figure can be viewed at wileyonlinelibrary.com]

stories showed some improvement from the pretest to a follow-up session, which was 4 months after the pretest session (a rise of 42% points in the feedback with explanation condition; a rise of 31% points in the feedback without explanation condition; a rise of 35% points in the no feedback condition; compared to a rise of 17% points in the control condition).

Similar to the fitted binomial linear mixed effect model that we presented in Table 3 (Model 2), in order to test the stability of the training effect, we fitted a binomial mixed effects model on the scores with an interaction between session (pretest and follow-up) and condition (feedback with explanation, feedback without explanation, no feedback, control); an interaction between condition and "three locations" scores at follow-up session; and story types ("three locations," "three goals," "decoy gift"), centered age, centered pretest working memory scores, and centered verbal ability score as fixed factors. As random effects, we had random slopes

for session per subject correlated with the random intercepts. The pretest session, "three goals" stories, and control condition were used as base levels in the model (reference categories). Table 4 shows the estimates and z-values of the binomial mixed-effects model for the stability of the training effect.

As can be seen from Table 4, in the control condition, children's second-order false belief scores did not significantly improve from pretest to follow-up sessions (Row 2). There was a significant difference from pretest to follow-up sessions between the control condition and the feedback with explanation condition (Row 12). Children in the feedback without explanation condition and the no feedback condition performed better than children in the control condition at the follow-up session; however, the differences in improvements between those conditions and the control condition were not significant (Rows 11, 13).

Moreover, children's improvements from pretest to follow-up sessions in "three locations" stories

Table 4

The Estimates and z-Values of the Binomial Mixed-Effects Model for the Stability of the Training Effect

		<i>B</i>	<i>SE</i>	<i>z</i>	<i>p</i>
1	(Intercept)	−1.32	.31	−4.26	< .001
2	Follow-up	0.35	.39	0.92	.36
3	Feedback without explanation	−0.71	.45	−1.58	.11
4	Feedback with explanation	−0.34	.42	−0.81	.42
5	No feedback	0.16	.38	0.41	.68
6	“Decoy gift”	1.64	.24	6.75	< .001
7	“Three locations”	−0.22	.35	−0.64	.53
8	Age	0.96	.34	2.78	.005
9	Verbal ability	0.002	.02	0.10	.92
10	Working memory	0.06	.10	0.58	.56
11	Follow-Up × Feedback Without Explanation	0.76	.62	1.22	.22
12	Follow-Up × Feedback With Explanation	1.47	.62	2.38	.02
13	Follow-Up × No Feedback	0.85	.58	1.48	.14
14	Control × “Three Locations” × Follow-Up	1.38	.57	2.41	.02
15	Follow-Up × Feedback Without Explanation × “Three Locations”	1.39	.70	2.01	.04
16	Follow-Up × Feedback With Explanation × “Three Locations”	0.92	.71	1.30	.19
17	Follow-Up × No Feedback × “Three Locations”	1.62	.68	2.40	.02

tended to be greater than children’s improvements in the trained story types (Rows 14, 15, 16, 17). However, unlike in the other conditions, in the feedback with explanation condition, children’s improvement from pretest to follow-up sessions on “three locations” did not significantly differ from the other story types (Row 16). In that condition, children’s improvements on the “three goals” stories were as great as on the “three locations” stories.

Finally, similar to the findings from pretest to posttest sessions that are shown in Table 3, although age had a significant effect, children’s verbal abilities and working memory scores at pretest did not have a significant effect on children’s second-order false belief score from pretest to follow-up sessions.

Discussion

For the first time in the literature, the roles of different types of feedback, namely feedback with

explanation, feedback without explanation, and no feedback have been studied in 5- to 6-year-olds’ development of second-order false belief reasoning. Crucially, the design of our study: (a) controls for the effects of age, working memory and verbal abilities; (b) tests children in a follow-up session after 4 months from the pretest session; (c) uses an active control condition; and (d) controls for children’s pattern recognition instead of reasoning about other people’s minds by testing them with second-order true belief stories.

Training Effects From Pretest to Posttest Sessions

As we predicted, children in the feedback with explanation condition made greater gains in second-order false belief reasoning from pretest to posttest than children in the active control group did, when the effects of age, verbal abilities, and working memory are controlled for. The positive training effect of feedback with explanation is in line with the previous findings of first-order false belief training studies (Appleton & Reddy, 1996; Clements et al., 2000; Melot & Angeard, 2003) and with the studies that tested 9- to 10-year-olds with more advanced ToM tasks in a group setting (Bianco et al., 2016; Lecce et al., 2014). This positive result of training 5- to 6-year-olds fills the gap in the ToM literature between preschool children and middle childhood.

Our second prediction, that there would also be a significant improvement in the feedback without explanation condition from pretest to posttest sessions compared to the control condition, was also confirmed. In this condition, the feedback “Correct” or “Wrong” together with the correct answer was provided without further explanation. However, as we predicted, children’s performance improved more when feedback with explanations had been provided, compared to providing feedback without explanations.

What do these improvements in feedback with and without explanation conditions mean in terms of children’s development of second-order false belief reasoning? A previous computational cognitive modeling study on second-order false belief reasoning provides an explanation. Arslan, Taatgen, et al.’s (2017) model contains task-independent declarative knowledge (*knowing that*): (a) the location of an object changes by an action toward that object; (b) seeing leads to knowing; (c) people search for objects at the location where they have last seen them unless they are informed that there is a change in the location of the object; (d) other

people reason “like me”. In addition, the model contains procedural knowledge for first-order and second-order false belief reasoning, for example, *knowing how* to apply false belief reasoning by retrieving relevant story facts to take the perspective of protagonists based on the task-independent declarative knowledge. Under the assumption that 5-year-olds have more experience using first-order than second-order false belief reasoning, the model initially uses first-order false belief reasoning when a second-order false belief question is asked (see Arslan, Taatgen, et al., 2017 for an empirical confirmation showing that most 5-year-olds use first-order false belief reasoning when they are asked second-order false belief questions). With repeated exposure to second-order false belief reasoning and with the help of the feedback “Wrong,” the model updates first-order false belief reasoning to second-order false belief reasoning and starts to give correct answers for second-order false belief questions. The model further predicts that when feedback with explanation is provided, the likelihood of revising the first-order false belief reasoning strategy to the correct second-order reasoning strategy will be higher, because explanations provide more evidence in favor of second-order false belief reasoning.

Based on the computational modeling study and in line with the studies on adults’ ToM reasoning in strategic games that we mentioned in the Introduction, it appears that the feedback “Wrong” together with the correct answer provides sufficient evidence for children that second-order false belief reasoning is needed. However, in line with the modeling study’s prediction, children’s performance improves more when the feedback “Wrong” is accompanied with explanations.

In addition to the feedback conditions, children’s performance from pretest to posttest sessions also increased in the no feedback condition (from 33% to 55% correct), although there was no significant difference between the no feedback and control conditions. To explain this trend in the no feedback condition, we surmise that exposing children to second-order false belief reasoning by asking them second-order false belief questions together with the justification questions “Why?” helps them to reflect about their own judgments (Amsterlaw & Wellman, 2006; Guajardo et al., 2013). Thus, asking justification questions helps them to revise their first-order false belief reasoning to correct second-order false belief reasoning. This argument needs to be tested with another training study in which children are trained on second-order false belief stories with no

feedback, however, this time without asking the justification questions. We predict that 5-year-olds’ second-order false belief reasoning cannot be significantly improved without the justification questions.

How can these findings be interpreted in terms of proposed theories of children’s development of second-order false belief reasoning? Our main assumptions were that 5-year-olds understand that they can use their first-order false belief reasoning recursively (i.e., conceptual change) and that they have sufficient language abilities to understand second-order false belief questions and the cognitive skills required to carry out second-order false belief reasoning without mistakes (i.e., complexity), however, they lack experience. The improvements in the feedback without explanation condition and the trend in the no feedback condition can be seen as evidence for these assumptions. Although we cannot rule out the possibility that our training program caused a conceptual change or helped children to overcome the complexity of the second-order false belief tasks; it is less likely that children understood for the first time that beliefs can be used recursively after hearing 12 second-order false belief stories, especially when they did not hear the explanations, and it is less likely that children’s language and executive functioning abilities were improved in the short time span of the training. Therefore, we propose that complexity and conceptual change are readiness factors for children to pass second-order false belief tasks; however, these factors are not sufficient to pass second-order false belief tasks. Children also need experience to realize that first-order false belief reasoning does not suffice to solve these second-order false belief tasks. Note that this does not mean that once children have experience they always perfectly apply second-order false belief reasoning. They might still make mistakes for different reasons, such as lack of efficiency in applying reasoning rules and internal or external distraction.

The role of experience in realizing that second-order false belief reasoning is needed can be interpreted as improvement in children’s metastrategic knowing, that is, metaknowing about procedural knowing (Kuhn, 2000; Kuhn & Pearsall, 1998). In line with Arslan, Taatgen, et al.’s (2017) computational modeling study, we argue that 5-year-olds have procedural knowledge (i.e., knowing how) to perform second-order false belief reasoning. However, due to lack of experience, they do not have explicit meta-level awareness that the second-order false belief reasoning strategy is needed and repeated exposure to second-order false belief

reasoning raises children's meta-level awareness for using second-order false belief reasoning (see Kuhn, 2000 for a similar argument about children's development in general).

How can our training study be interpreted in terms of children's real-life experiences? Children do not repeatedly encounter many second-order false belief stories and get immediate feedback in their daily life. However, previous research has shown that social factors, such as parents' story-book reading (de Rosnay & Hughes, 2006), conversation about mental states (Heyes & Frith, 2014; Peterson & Slaughter, 2003; Ruffman, Slade, & Crowe, 2002) and providing reasons when correcting children's misbehavior (Ruffman, Perner, & Parkin, 1999) all contribute to children's development of ToM. It appears likely that these social factors are also the source of gaining experiences in second-order false belief reasoning, as witnessed by the great popularity among 5-year-olds of the picture book *"The Gruffalo"* (Donaldson & Scheffler, 1999). The excitement of the story hinges on the fact that a mouse successfully makes the large and scary Gruffalo falsely *believe* that all the other animals in the forest *believe* that the mouse is very dangerous. To validate one of our assumptions, namely, that 5-year-olds have less experience with second-order ToM reasoning than first-order ToM reasoning, it would be worthwhile to investigate children's everyday life experiences with a corpus study.

Generalizability of the Training Effect

It is also worth discussing the differences between the three story types that we used in our study, namely "decoy gift," "three goals," and "three locations." "Decoy gift" stories are less complex than the "three goals" and "three locations" stories because "decoy gift" stories have two possible answers (50% chance level) compared to three possible answers (33% chance level) in the other story types. In line with this explanation, in general, children's scores were higher in "decoy gift" stories. On the other hand, "three goals" stories and "three locations" stories both have three possible answers for the second-order false belief question; Indeed, there was no significant difference in children's scores between "three goals" and "three locations" stories at the pretest session.

Note that we chose not to use "three locations" stories during the training sessions in order to test the generalizability of the training effect. In the experimental conditions, children's improvements in the stories that they trained on (i.e., "three goals"

and "decoy gift" stories) did not significantly differ from their improvements in "three locations" stories from pretest to posttest sessions (see Figure 5 and Table 3). These results together show the generalizability of the training effect.

Moreover, as can be seen in Table 4, we found an interesting and unexpected result about children's improvements in "three locations" stories from pretest to follow-up sessions. In all conditions, children's improvements from pretest to follow-up sessions in "three locations" stories were greater than their improvements in the stories that we used in the training sessions. More improvement in "three locations" over 4 months might be related to the linguistic structural differences in the second-order false belief questions in "three locations" stories compared to the second-order false belief questions in "decoy gift" and "three goals" stories. In "three locations" stories, the structure of the second-order false belief questions was in the form of a second-order embedding (e.g., "Where does Marieke think that Kevin will look for the chocolate?"). On the other hand, in "decoy gift" and "three goals" stories, the second-order false belief questions did not involve second-order embedding and were broken down into two pieces in order to facilitate children's comprehension (e.g., "The grandma asks dad: 'What does Robert think that you will do today?' followed by the second-order false belief question 'What does dad say to grandma?'").

However, once children are more competent with second-order embedding, a second-order false belief question in the form of a second-order embedding might facilitate reasoning by delivering proper chunks ready for serialization (Arslan, Hohenberger, et al., 2017; de Villiers et al., 2014; Hollebrandse & Roeper, 2014). Further research is needed to investigate children's development over time in answering those two types of questions. Alternatively, in order to see whether children's improvements in the "decoy gift" and "three goals" stories will be as great as their improvements in the "three locations" stories at the follow-up session, our training program can be replicated by asking the second-order false belief questions in second-order embedding form for all types of stories at the follow-up session.

Stability of the Training Effects: Improvements From Pretest to Follow-Up Sessions

As can be seen from Figure 5 and Table 4, for all the conditions, children who were in one of the three experimental conditions performed better at the follow-up session than the children who were in

the control condition (73% correct in the feedback with explanation condition; 56% in the feedback without explanation condition; 68% in the no feedback condition; compared to 46% in the control condition). However, the greatest improvement from pretest to follow-up session occurred in the feedback with explanation group, which significantly differed from the control condition. This result emphasizes the importance of further explanations in children's development of second-order ToM and is in line with Slaughter and Gopnik's (1996) theory, which suggests that experiences relevant to the domain of intuitive psychology takes time to influence a child's theoretical conceptual structure. The small improvement from pretest to follow-up sessions in the control condition can be interpreted as the effect of children's natural development over 4 months combined with the effect of exposure to the second-order false belief tasks nine times (three stories each at pretest, posttest, and follow-up sessions).

Limitations

In addition to the above-mentioned strengths of the design and the novelty of our findings, our study also has some limitations. First, in line with the previous training studies of ToM, each child was tested by the same experimenter across the training sessions (Appleton & Reddy, 1996; Bianco et al., 2016; Cavallini et al., 2015; Clements et al., 2000; Hale & Tager-Flusberg, 2003). However, use of the same experimenter for all of the sessions might cause experimenter-bias and good-subject effects (Nichols & Edlund, 2015). Moreover, the children in the feedback with and without explanation conditions were older than those in the no feedback and control conditions. Although we did further statistical analyses to account for these differences, it would be worthwhile to replicate our experiment in the future, by assigning different experimenters for each session in order to investigate possible experimenter-bias and good-subject effects as well as using random assignment with constraints to ensure that groups end up equivalent on the variable of interests, such as age, pretest performance and executive function.

Second, to assess the generalizability of our training program, we used a second-order false belief story in the pretest, posttest, and follow-up of a type on which children were not trained in the training sessions. Considering the previous correlational studies showing that second-order false belief reasoning is associated with more advance social skills, it would be worthwhile to investigate whether

children can generalize what they have learned from the second-order false belief training to their everyday social competence, such as idiom and irony understanding, maintaining lies and catching liars: "ToM in action" (cf. Talwar & Lee, 2008). Moreover, in order to investigate the role of experience in second-order ToM reasoning similar to real life, designing a training study in which 5-year-olds are trained with narrative fictions and games that require second-order ToM is worthwhile.

Finally, we did not find any effect of children's verbal abilities and working memory score at pretest on children's second-order false belief score. The absence of a significant effect of working memory is in line with Hasselhorn et al.'s (2005) study. We interpret the lack of a significant effect of working memory on children's second-order false belief score as a result of the simplicity of the counting span task. In order to succeed in the counting span task, children need only count the blue shapes and make a list of numbers in their memory to report later. In contrast, we believe that a working memory task that requires more complex working memory strategies might predict children's second-order false belief scores at pretest. For example, a listening span task in which children are expected to first judge the truthfulness of each sentence by saying "yes" or "no" and then have to recall the last word of all the sentences of a set told to them so far, in reverse order, can be a predictor of children's second-order false belief reasoning development (see Arslan, Hohenberger, et al., 2017). Further research is needed to verify this prediction.

These limitations aside, the relatively high proportion of correct judgments that were adequately explained can be seen as evidence for the validity of our training program. We believe our study has important findings both in terms of theories of second-order false belief reasoning and its real-life application. For the first time in the literature, it has been shown that 5-year-olds can be helped to cross the threshold to second-order false belief reasoning by the exposure to many stories and by asking them to reflect on second-order false belief questions, without providing any explanations about their wrong answers. Based on our findings, we propose that 5-year-olds' failures in second-order false belief tasks are due to lack of *experience*.

References

- Amsterlaw, J., & Wellman, H. M. (2006). Theories of mind in transition: A microgenetic study of the

- development of false belief understanding. *Journal of Cognition and Development*, 7, 139–172. https://doi.org/10.1207/s15327647jcd0702_1
- Appleton, M., & Reddy, V. (1996). Teaching three year-olds to pass false belief tests: A conversational approach. *Social Development*, 5, 275–291. <https://doi.org/10.1111/j.1467-9507.1996.tb00086.x>
- Arslan, B., Hohenberger, A., & Verbrugge, R. (2017a). Syntactic recursion facilitates and working memory predicts recursive theory of mind. *PLoS ONE*, 12, e0169510. <https://doi.org/10.1371/journal.pone.0169510>
- Arslan, B., Taatgen, N., & Verbrugge, R. (2017b). Five-year-olds' systematic errors in second-order false belief tasks are due to first-order theory of mind strategy selection: A computational modeling study. *Frontiers in Psychology*, 8, 275. <https://doi.org/10.3389/fpsyg.2017.00275>
- Astington, J. W., Pelletier, J., & Homer, B. (2002). Theory of mind and epistemological development: The relation between children's second-order false-belief understanding and their ability to reason about evidence. *New Ideas in Psychology*, 20, 131–144. [https://doi.org/10.1016/S0732-118X\(02\)00005-3](https://doi.org/10.1016/S0732-118X(02)00005-3)
- Bianco, F., Lecce, S., & Banerjee, R. (2016). Conversations about mental states and theory of mind development during middle childhood: A training study. *Journal of Experimental Child Psychology*, 149, 41–61. <https://doi.org/10.1016/j.jecp.2015.11.006>
- Caillies, S., & Le Sourn-Bissaoui, S. (2013). Nondecomposable idiom understanding in children: Recursive theory of mind and working memory. *Canadian Journal of Experimental Psychology/Revue Canadienne de Psychologie Expérimentale*, 67, 108–111. <https://doi.org/10.1037/a0028606>
- Cavallini, E., Bianco, F., Bottiroli, S., Rosi, A., Vecchi, T., & Lecce, S. (2015). Training for generalization in Theory of Mind: A study with older adults. *Frontiers in Psychology*, 6, 1123.
- Clements, W., Rustin, C. L., & McCallum, S. (2000). Promoting the transition from implicit to explicit understanding: A training study of false belief. *Developmental Science*, 3, 81–92. <https://doi.org/10.1111/1467-7687.00102>
- de Rosnay, M., & Hughes, C. (2006). Conversation and theory of mind: Do children talk their way to socio-cognitive understanding? *British Journal of Developmental Psychology*, 24, 7–37. <https://doi.org/10.1348/026151005X82901>
- de Villiers, J., Hobbs, K., & Hollebrandse, B. (2014). Recursive complements and propositional attitudes. In M. Speas & T. Roeper (Eds.), *Recursion: Complexity in cognition. Theoretical studies in psycholinguistics*, (Vol. 43, pp. 221–242). Heidelberg, Germany; New York, NY: Springer.
- Donaldson, J., & Scheffler, A. (1999). *The gruffalo*. Basingstoke, UK: Pan MacMillan.
- Filippova, E., & Astington, J. W. (2008). Further development in social reasoning revealed in discourse irony understanding. *Child Development*, 79, 126–138. <https://doi.org/10.1111/j.1467-8624.2007.01115.x>
- Flobbe, L., Verbrugge, R., Hendriks, P., & Krämer, I. (2008). Children's application of theory of mind in reasoning and language. *Journal of Logic, Language and Information*, 17, 417–442. <https://doi.org/10.1007/s10849-008-9064-7>
- Goodman, N. D., Baker, C. L., Bonawitz, E. B., Mansinghka, V. K., Gopnik, A., Wellman, H., . . . & Tenenbaum, J. B. (2006). Intuitive theories of mind: A rational approach to false belief. In R. Sun & N. Miyake (Eds.), *Proceedings of the twenty-eighth annual conference of the Cognitive Science Society* (pp. 1382–1390). Vancouver, BC: Cognitive Science Society.
- Guajardo, N. R., Petersen, R., & Marshall, T. R. (2013). The roles of explanation and feedback in false belief understanding: A microgenetic analysis. *Journal of Genetic Psychology*, 174, 225–252. <https://doi.org/10.1080/00221325.2012.682101>
- Hale, C. M., & Tager-Flusberg, H. (2003). The influence of language on theory of mind: A training study. *Developmental Science*, 6, 346–359. <https://doi.org/10.1111/1467-7687.00289>
- Happé, F. (1994). An advanced test of theory of mind: Understanding of story characters' thoughts and feelings by able autistic, mentally handicapped, and normal children and adults. *Journal of Autism and Developmental Disorders*, 24, 129–154. <https://doi.org/10.1007/BF02172093>
- Hasselhorn, M., Mähler, C., & Grube, D. (2005). Theory of mind, working memory, and verbal ability in preschool children: The proposal of a relay race model of the developmental dependencies. In W. Schneider, R. Schumann-Hengsteler, & B. Sodian (Eds.), *Young children's cognitive development: Interrelations among executive functioning, working memory, verbal ability, and theory of mind* (pp. 219–237). Mahwah, NJ: Erlbaum.
- Heyes, C. M., & Frith, C. D. (2014). The cultural evolution of mind reading. *Science*, 344, 1243091. <https://doi.org/10.1126/science.1243091>
- Hofmann, S. G., Doan, S. N., Sprung, M., Wilson, A., Ebesutani, C., Andrews, L. A., . . . & Harris, P. L. (2016). Training children's theory-of-mind: A meta-analysis of controlled studies. *Cognition*, 150, 200–212. <https://doi.org/10.1016/j.cognition.2016.01.006>
- Hollebrandse, B., Hobbs, K., de Villiers, J., & Roeper, T. (2008). Second order embedding and second order false belief. In A. Gavarro & J. M. Freitas (Eds.), *Proceedings of generative approaches to language acquisition (GALA)* (pp. 270–280). Newcastle, UK: Cambridge Scholar Press.
- Hollebrandse, B., & Roeper, T. (2014). Empirical results and formal approaches to recursion in acquisition. In M. Speas & T. Roeper (Eds.), *Recursion: Complexity in cognition theoretical studies in psycholinguistics* (pp. 179–219). Heidelberg, NY: Springer.
- Hollebrandse, B., van Hout, A., & Hendriks, P. (2014). Children's first and second-order false-belief reasoning in a verbal and a low-verbal task. *Synthese*, 191, 321–333. <https://doi.org/10.1007/s11229-012-0169-9>

- Kloo, D., & Perner, J. (2003). Training transfer between card sorting and false belief understanding: Helping children apply conflicting descriptions. *Child Development, 74*, 1823–1839. <https://doi.org/10.1046/j.1467-8624.2003.00640.x>
- Kloo, D., & Perner, J. (2008). Training theory of mind and executive control: A tool for improving school achievement? *Mind, Brain, and Education, 2*, 122–127. <https://doi.org/10.1111/j.1751-228X.2008.00042.x>
- Kuhn, D. (2000). Metacognitive development. *Current Directions in Psychological Science, 9*, 178–181. <https://doi.org/10.1111/1467-8721.00088>
- Kuhn, D., & Pearsall, S. (1998). Relations between metacognitive knowledge and strategic performance. *Cognitive Development, 13*, 227–247. [https://doi.org/10.1016/S0885-2014\(98\)90040-5](https://doi.org/10.1016/S0885-2014(98)90040-5)
- Lecce, S., Bianco, F., Devine, R. T., & Hughes, C. (2014). Promoting theory of mind during childhood: A training program. *Journal of Experimental Psychology, 126*, 52–67. <https://doi.org/10.1016/j.jecp.2014.03.002>
- Leslie, A. M. (1987). Pretence and representation: The origins of “theory of mind.” *Psychological Review, 94*, 412–426. <https://doi.org/10.1037/0033-295X.94.4.412>
- Lohmann, H., & Tomasello, M. (2003). The role of language in the development of false belief understanding: A training study. *Child Development, 74*, 1130–1144. <https://doi.org/10.1111/1467-8624.00597>
- Meijering, B., Taatgen, N. A., Van Rijn, H., & Verbrugge, R. (2014). Modeling inference of mental states: As simple as possible, as complex as necessary. *Interaction Studies, 15*, 455–477. <https://doi.org/10.1075/is.15.3.05mei>
- Melot, A. N., & Angeard, N. (2003). Theory of mind: Is training contagious? *Developmental Science, 6*, 178–184. <https://doi.org/10.1111/1467-7687.00269>
- Miller, S. A. (2009). Children’s understanding of second-order mental states. *Psychological Bulletin, 135*, 749–773. <https://doi.org/10.1037/a0016854>
- Miller, S. A. (2012). *Theory of mind: Beyond the preschool years*. Hove, UK: Psychology Press. <https://doi.org/10.4324/97802031222730>
- Nichols, A. L., & Edlund, J. E. (2015). Practicing what we preach (and sometimes study): Methodological issues in experimental laboratory research. *Review of General Psychology, 19*, 191–202. <https://doi.org/10.1037/gpr0000027>
- Ornaghi, V., Brockmeier, J., & Gavazzi, I. G. (2011). The role of language games in children’s understanding of mental states: A training study. *Journal of Cognition and Development, 12*, 239–259. <https://doi.org/10.1080/15248372.2011.563487>
- Perner, J., Kain, W., & Barchfeld, P. (2002). Executive control and higher-order theory of mind in children at risk of ADHD. *Infant and Child Development, 11*, 141–158. [https://doi.org/10.1002/\(ISSN\)1522-7219](https://doi.org/10.1002/(ISSN)1522-7219)
- Perner, J., & Wimmer, H. (1985). “John thinks that Mary thinks that. . .”: Attribution of second-order beliefs by 5- to 10-year old children. *Journal of Experimental Child Psychology, 5*, 125–137. [https://doi.org/10.1016/0022-0965\(85\)90051-7](https://doi.org/10.1016/0022-0965(85)90051-7)
- Peterson, C., & Slaughter, V. (2003). Opening windows into the mind: Mothers’ preferences for mental state explanations and children’s theory of mind. *Cognitive Development, 18*, 399–429. [https://doi.org/10.1016/S0885-2014\(03\)00041-8](https://doi.org/10.1016/S0885-2014(03)00041-8)
- Premack, D., & Woodruff, G. (1978). Does the chimpanzee have a theory of mind? *Behavioral and Brain Sciences, 4*, 515–526. <https://doi.org/10.1017/S0140525X00076512>
- Ruffman, T., Perner, J., & Parkin, L. (1999). How parenting style affects false belief understanding. *Social Development, 8*, 395–411. <https://doi.org/10.1111/1467-9507.00103>
- Ruffman, T., Slade, L., & Crowe, E. (2002). The relation between children’s and mothers’ mental state language and theory-of-mind understanding. *Child Development, 73*, 734–751. <https://doi.org/10.1111/1467-8624.00435>
- Schmidt, A. M. G., & Westendorp, F. (2011). *Jip en Janneke* [Jip and Janneke]. Amsterdam, The Netherlands: Querido.
- Slaughter, V. (1998). Children’s understanding of pictorial and mental representations. *Child Development, 69*, 321–332. <https://doi.org/10.1111/j.1467-8624.1998.tb06191.x>
- Slaughter, V., & Gopnik, A. (1996). Conceptual coherence in the child’s theory of mind: Training children to understand belief. *Child Development, 67*, 2967–2988. <https://doi.org/10.2307/1131762>
- Sodian, B., Taylor, C., Harris, P., & Perner, J. (1992). Early deception and the child’s theory of mind: False trails and genuine markers. *Child Development, 62*, 468–483. <https://doi.org/10.1111/j.1467-8624.1991.tb01545.x>
- Sullivan, K., Zaitchik, D., & Tager-Flusberg, H. (1994). Preschoolers can attribute second-order beliefs. *Developmental Psychology, 30*, 395–402. <https://doi.org/10.1037/0012-1649.30.3.395>
- Talwar, V., & Lee, K. (2008). Social and cognitive correlates of children’s lying behaviour. *Child Development, 79*, 866–881. <https://doi.org/10.1111/j.1467-8624.2008.01164.x>
- Towse, J. N., Hitch, G. J., & Hutton, U. (1998). A reevaluation of working memory capacity in children. *Journal of Memory and Language, 39*, 195–217. <https://doi.org/10.1006/jmla.1998.2574>
- Wellman, H. M. (1990). *The child’s theory of mind*. Cambridge, MA: MIT Press.
- Wellman, H. M., Cross, D., & Watson, J. (2001). Meta-analysis of theory-of-mind development: The truth about false belief. *Child Development, 72*, 655–684. <https://doi.org/10.1111/1467-8624.00304>
- Wimmer, H., & Perner, J. (1983). Beliefs about beliefs: Representation and constraining function of wrong beliefs in young children’s understanding of deception. *Cognition, 13*, 103–128. [https://doi.org/10.1016/0010-0277\(83\)90004-5](https://doi.org/10.1016/0010-0277(83)90004-5)

Appendix A

Non-Experimental Questions

Table A1 shows the percentages of correct answers (standard errors) for the non-experimental questions. Control questions and first-order false belief questions were asked before the second-order false belief questions, in order to make sure that children did not have problems with memorizing the story facts and understanding the relatively complex second-order false belief questions as well as attributing first-order false beliefs. Moreover, in order to make sure that children did not use a pattern learning strategy instead of reasoning about the second-order false belief questions, we investigated children's performance on second-order true belief questions.

Appendix B

In line with the previous literature (Perner & Wimmer, 1985; Sullivan et al., 1994), we considered a justification answer as correct if it is appropriate and belongs to one of the following types:

Explicit Second-Order Reasoning

The child embeds one character's epistemic state in the other character's mental state, for example, "Because she believes that Kevin doesn't know that the chocolate is in the toy box."

Table A1
The Percentages of Correct Answers (Standard Errors) for the Non-Experimental Questions

Question type	Total number of questions	Correct answers
Control questions	4,968	98% (0.002)
First-order false belief	1,811	95% (0.005)
First-order true belief	286	97% (0.01)
Second-order true belief	286	92% (0.02)
Questions in neutral stories in the control condition	1,900	95% (.005)

As can be seen in Table A1, children predominantly gave correct answers to these non experimental questions, meaning that they did not have problems with memorizing the story facts, nor with first-order false belief reasoning. The high proportion of correct second-order true belief answers (different from the corresponding second-order false belief answers) shows that children's improvements in second-order false belief tasks cannot be attributed to pattern learning instead of applying second-order theory of mind reasoning.

Implicit Second-Order Reasoning

Relevant information is embedded in one character's epistemic state, for example, "Because she doesn't know that Kevin saw it."

Perceptive Information

Relevant information is embedded in one character's perception, for example, "Because she didn't see that Kevin was looking through the window."

Communicative Information

Information is mentioned that was communicated to the secondary character, for example, "Because she said she bought a ball."

Location Information

The original location of the object is mentioned, for example, "Because Kevin put it into the drawer before."

Appendix C

An Example Neutral Story: Jip's Got a Ponytail

(translation by David Colman from the original Dutch version by Annie M.G. Schmidt)

Jip is at the barber's. Snip, snip, go the scissors. And Jip says, 'Ow!'

"I'm not hurting you," says the barber. "How old are you anyway? I didn't even touch you."

Snip, snip, go the scissors. And Jip hates it. He just hates it. He keeps shouting, "Ow, ow!"

"Just a little bit more," says the barber.

But Jip shouts "Ow!" one more time. He jumps up and runs out of the shop wearing the white barber's cape.

"Hey, where are you going?" the barber shouts. "You're not finished yet! Just a little longer!"

But Jip has had enough. He runs very fast. And the barber runs very fast after him. But Jip is faster. He's almost home and the barber shrugs and gives up.

Jip sits down by the side of the road. He is still wearing the white barber's cape. He's crying because he was so scared.

Here comes Janneke. She sees Jip sitting there. And she starts laughing. She can't help it. "You look really silly," she says.

Jip looks up and stops crying.

"You've got a ponytail on top of your head," says Janneke. And she snorts with laughter. "You've got a ponytail and you're wearing a serviette."

And it's true. Jip is almost completely bald, but there is a tuft of hair left on the top of his head. Just like a little ponytail.

At this point the following two questions were asked to the children: "Where is Jip at the beginning of the story?" and "What is the reason why Janneke laughs about Jip?"

After children gave answers for these neutral questions, the story continued as follows:

Janneke laughs so much it makes Jip angry. "I'm not going back to the barber's," he says.

"Then you'll have to walk around with a ponytail for the rest of your life," Janneke says. "Jip's got a ponytail, Jip's got a ponytail!"

That really is horrible. Having Janneke laugh at him! That's too much. Jip gets up very slowly and very slowly he walks back to the barber's. "The ponytail has to come off," he says.

"I told you that already!" grumbles the barber. "You ran off before I was finished." And Jip has to get back on the chair.

Snip, snip, go the scissors.

"There," says the barber. "Now it's done. Was it really that bad?"

Jip smiles. The barber undoes the white cape, and Jip is back out on the street.

"My ponytail's gone," he says to Janneke. "I can see that," says Janneke. "It looked really, really silly."

And then they play marbles.

At the end of the story, the following two questions were asked: "Does Jip go back to the barber?" and "Does the barber leave the ponytail at the second time, too?"

Appendix D

Figure D1 shows the comparison of the younger children's improvements (age range = 4;8–5;6) from pretest to posttest sessions in the experimental conditions ($N = 35$, $M_{\text{age}} = 5;3$, $SD = 0.2$) with the younger children in the active control condition ($N = 27$, $M_{\text{age}} = 5;3$, $SD = 0.2$).

Figure D1. Proportion of correct answers to the second-order false belief questions for the younger children in the experimental conditions in comparison with the younger children in the control condition. [Color figure can be viewed at wileyonlinelibrary.com]