A computational cognitive modeling approach to the development of second-order theory of mind
Arslan, Burcu

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:
2017

Link to publication in University of Groningen/UMCG research database

Citation for published version (APA):
Chapter 3:
Five-year-old Children’s Development of Second-order False Belief Reasoning Can Be Accelerated: An Evaluation of Different Feedback Methods

In which we investigate the role of different types of feedback on children’s transitions from failure to success in second-order false belief tasks.

This chapter is in preparation to submit to a journal and a short version was previously published as:
Abstract

For the first time in the literature, we have conducted a training study in order to accelerate 5- to 6-year-olds’ development of second-order false belief reasoning. Previous studies have shown that it is only possible to accelerate 4-year-olds’ development of first-order theory of mind by providing feedback with explanations but not feedback without explanations or without any feedback. Following these findings, recent studies used feedback with explanations to accelerate 9-year-olds’ development of advanced theory of mind. In order to assess whether the findings of the first-order theory of mind literature still hold, we trained 106 children with second-order false belief tasks in one of the following conditions: (i) Feedback with explanation; (ii) Feedback without explanations; (iii) No feedback; (iv) Active control. The results showed that there were significant improvements in children’s scores from pre-test to post-test in the three experimental conditions, from 31% to 68% in the feedback with explanation condition: from 25% to 49% in the feedback without explanation condition; and from 33% to 55% in the no feedback condition, compared to a small improvement in the active control condition (from 29% to 35%). Moreover, the improvements were not due to children’s age, verbal abilities and working memory scores. Importantly, the children were able to generalize the training effect to another story type that they had not been trained on, and the training effect was stable at a follow-up session 4 months after the pre-test. Overall, our results highlight the difference between first-order and second-order theory of mind development and suggest that children can be helped over the threshold to second-order false belief reasoning by exposure to many stories and by asking them to reflect on second-order false belief questions, without providing explanations about their wrong answers.

Keywords: theory of mind; development; training study; second-order false belief reasoning; feedback

3.1. Introduction

In daily life, children are constantly in interaction with their friends and family members and 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). It has been argued that being able to attribute a false belief to someone else provides evidence that a person has a theory of mind (Dennett, 1978; Wellman, 1990). Since then, false belief tasks have become some of the most applied tasks for testing children’s development of ToM. A first-order false belief task examines whether children can attribute a false belief to a protagonist in a given story where the child knows the reality, while the protagonist has a false belief about the reality. Many studies have shown that, before the age of 4, most children cannot pass verbal first-order false belief tasks (Wimmer & Perner, 1983; 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 mental state to another character in the story (Perner & Wimmer, 1985; Sullivan et al., 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. While first-order false belief reasoning is found to be related to social skills, such as deception (Sodian, Taylor, Harris, & Perner, 1992; Bosco & Gabbatore, 2017) and pretend play (Leslie, 1987), second-order false belief reasoning is found to be important in more advanced aspects of children’s everyday social competence, such as idiomatic understanding (Caillies & Le Sourn-Bissaoui, 2015), irony understanding (Filippova & Astington, 2008; Bosco & Gabbatore, 2017), and reasoning about evidence (Astoning, Pelletier, & Homer, 2002). As a concrete example, to successfully maintain a strategic lie, the liar has to reason about what the listener knows about what the liar knows, requiring second-order theory of mind (Talwar & Lee, 2008). Considering the importance of second-order ToM in children’s more advanced social skills, it is important to find effective methods to accelerate children’s development as well as to understand children’s development of second-order ToM.

3.1.1. Training studies on first-order theory of mind: Related work

Several training studies have shown that it is possible to accelerate pre-school children’s development of first-order ToM with a moderately strong effect size
Chapter 3: Five-year-old children’s development of second-order false belief reasoning can be accelerated

3.1. Introduction

ToM reasoning or with tasks that are thought to be irrelevant for ToM reasoning. ToM tasks (Appleton & Reddy, 1996; Clements, Rustin, & McCallum, 2000; Lecce, Bianco, Demichelli, & Cavallini, 2014a; Melot & Angeard, 2003; Slaughter & Gopnik, 1996) or with different cognitive tasks that correspond to influential factors in children’s ToM development, such as language tasks (Hale & Tager-Flusberg, 2003; Lohmann & Tomasello, 2003; Shuliang, Yanjie, Sabbagh, & Bianco, Demichelli, & Cavallini, 2014a; Melot & Angeard, 2003; see also Kloo & Perner, 2008 for a review). As an active control condition, children were trained either with stories that did not involve ToM reasoning or with tasks that are thought to be irrelevant for ToM reasoning. Generally, one to two weeks after the training sessions, children were post-tested to assess whether they had improved significantly since their initial performance at pre-test, compared to the control condition.

Importantly, most of the studies that trained children with ToM tasks used corrective feedback by providing further explanations during the training sessions. Unfortunately, almost none of the studies tested children again in a follow-up session a couple of months after the post-test session in order to assess whether children’s improvements were stable over time (Hoffman et al., 2016). For positive exceptions, see Bianco, Lecce, & Banerjee (2016), Lecce et al. (2014a), and Lecce et al. (2014b), who tested children two months after the post-test. In our study on second-order ToM, we investigate different types of feedback to accelerate children’s development of second-order false belief reasoning, and we do test children in a follow-up session four months after the pre-test session, in order to see whether their improvements are stable.

Two first-order ToM studies that trained children by providing feedback without explanations or without feedback are noteworthy for our study. Clements et al. (2000) tested 91 children between the ages 3 and 5 with first-order false belief tasks in four sessions, each 7 days apart (i.e., pre-test, training day 1, training day 2, post-test). In both the experimental and practice conditions, children heard one false belief story per day. While in the experimental condition children received feedback with detailed explanations during the training sessions, in the practice condition they only received the feedback “Correct/Wrong”, without further explanation. In the control condition, children listened to neutral stories that did not involve any ToM reasoning. While children’s scores in the experimental condition significantly improved from pre-test to post-test, children who were in the practice and control conditions did not show significant improvements.

Similarly, Melot and Angeard (2003) trained 93 children between the ages 4 and 5 with first-order ToM tasks. At both pre-test and post-test, children were evaluated on first-order false belief tasks and appearance-reality tasks (e.g., an imitation pencil made out of rubber). Children who did not succeed both the false belief tasks and the appearance-reality tasks were trained with only six first-order false belief tasks in one experimental condition, and with only six appearance-reality distinction tasks in a second experimental condition. In both experimental conditions, feedback together with explanations was provided during the training sessions. Children in the control group were trained with three first-order false belief tasks in one training session, and with three appearance-reality tasks in another training session. Different from the experimental conditions, children did not get any feedback in the control condition. The results showed that children’s performance at the post-test had improved in comparisons to the pre-test in both of the experimental conditions but not in the control condition in which children were tested with first-order ToM tasks but did not get any feedback.

In summary, Clements et al.’s (2000) and Melot and Angeard’s (2003) studies have emphasized the importance of explicit feedback with detailed explanations in 3- to 5-year-old children’s improvements in first-order ToM, and have shown that children’s first-order ToM performance could not be improved by providing feedback without explanations or without providing any feedback. Based on these results, as we mentioned above, almost all of the previous training studies of ToM have provided feedback with detailed explanations in the training sessions in their experimental conditions.

3.1.2. Training studies on second-order theory of mind: Related work

Given that children’s ToM development goes beyond first-order false belief reasoning and continues to develop after they reach the age of 5, it is important to know whether it also holds that training children with second-order false belief tasks is only useful when feedback with explanations are provided. Moreover, understanding which types of feedback accelerate children’s second-order ToM might help to understand the underlying mechanism of children’s ToM development beyond the pre-school years (see Miller, 2009; 2012 for an extensive review).

However, as far as we know, there are no training studies on second-order false belief reasoning yet. There are two important previous training studies on children’s ToM development beyond the pre-school years. Lecce et al. (2014b) designed a conversation-based training study for 9- to 10-year-old children to
investigate the efficacy of the conversations about mental states in children’s development of advanced ToM. Because most children around the age of 9 already pass second-order false tasks, Lecce et al. used a more advanced and naturalistic ToM task – an Italian version of the Strange Stories task (Happé, 1994) – in which children’s ability to make inferences about mental states in nonliteral statements was assessed. During the training sessions, children participated in a group conversation about the stories and got corrective feedback and further explanations. In the control condition, children had to reason about similar stories to the Strange Stories task, however, this time, the stories involved physical events instead of mental states reasoning. Their findings showed that children’s performance from pre-test to post-test significantly improved for children in the experimental condition compared to the children in the control condition, and this improvement was stable over 2 months. Following Lecce et al.’s (2014b) findings, Bianco et al. (2016) focused on the question if and how conversations about mental states contribute 9- to 10-year-old children’s development of advanced ToM. They tested two possible explanations, namely frequent use of mental-state lexicon and accuracy of mental-state attribution. They applied the same training procedure of Lecce et al.’s (2014b) study where children participated in a group discussion and got corrective feedback with further explanations. They concluded that the accuracy of mental-state attribution, but not the frequency of mental-state lexicon, mediated the positive effect of conversations about mental states on children’s advanced ToM development.

In summary, both of those studies support the previous findings from the first-order ToM literature about the positive role of feedback with explanations for children’s further development of ToM. However, in those two studies children did not train with a condition in which they would do advanced ToM tasks but would only get feedback “Correct/Wrong” together with the correct answer without further explanations, or would get no 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 ToM tasks for children older than 4 but younger than 9, especially those children on the brink of developing second-order ToM.

3.1.3. The current study

In this study, we aim to fill 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 12 different second-order false belief stories with different types of feedback: (i) Feedback with explanation: by providing feedback “Correct/Wrong” together with the correct answer and further explanations about the reason why it is the answer; (ii) Feedback without explanations: by providing feedback “Correct/Wrong” together with the correct answer but without further explanations; (iii) No feedback. To the best of our knowledge, this is both the first time that children have been tested with such a large number of different second-order false belief stories, and the first time that the roles of different types of feedback on second-order false belief reasoning have been investigated.

We have two specific hypotheses about the possible effects of training children using feedback with explanation and feedback without further explanations. 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 pre-test to post-test sessions, and that they will improve significantly more than the children who are in the control condition.

Our second hypothesis is based on a previous computational modeling study that simulated children’s transitions from first-order ToM to second-order ToM (Arslan, Taatgen, & Verbrugge, 2013; 2017a). Contrary to the first-order ToM literature, one of the predictions of the computational cognitive model was that 5-year-old children who mastered first-order ToM can, in principle, pass second-order false belief tasks with the help of the feedback “Correct/Wrong” without any need for further explanation. Arslan et al. (2013; 2017a) have argued that the problem that most 5-year-old children who cannot pass second-order false belief tasks still encounter is that children are not used to reasoning about second-order mental states, therefore, they are not able to select the correct second-order ToM strategy to answer a second-order false belief question (e.g., “Where does Ayla think that Murat will look for the chocolate?”). They proposed that children can revise their wrong ToM reasoning strategy (i.e., first-order ToM) and can pass second-order false belief tasks by getting sufficient exposure to second-order false belief reasoning and getting the feedback “Correct/Wrong” without any need of further explanation. Therefore, considering the prediction of the computational modeling study, we expect that children who are in the feedback without explanation condition will also show an improvement from pre-test to post-test sessions.

In addition to second-order false belief tasks, we tested children with a working memory task (see Method section for the details). The rationale for testing children with a working memory task was based on the findings of previous literature on the role of executive functions in children’s development of first-order ToM (Benson, Sabbagh, Carlson, & Zelazo, 2013; Carlson, Moses, & Breton, 2002; Carlson, Moses, & Claxton, 2004; Devine & Hughes, 2014), and second-order ToM (Perner, Kain, & Barchfeld, 2002; but see Hasselhorn, Mählner, & Grube, 2005 for no significant correlation between the working memory span score and
3.2. Method

3.2.1. Participants

One hundred and nineteen children were recruited from a primary school with predominantly upper-middle-class families from Groningen, the Netherlands. All children had Dutch as their first language, and they were students of three different teachers. All different conditions contained about equal numbers from all these three teachers who teach 5- to 6-year-olds. The analysis showed that there was no significant effect of the teacher in children’s scores at pre-test and improvements after the training sessions, and adding teachers as a random effect did not improve the linear mixed effect models. Therefore, we merged the data across the teachers for the rest of the analysis. We sent a written parental consent 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. Children were pre-tested 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 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 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 pre-test. Thus, the analysis included the results of 106 children in three experimental conditions and one control condition. Table 3.1 illustrates the number of participants (the number of female participants in parentheses), the age range, and the mean age (standard deviations in parentheses), the verbal ability scores (standard errors in parentheses), the working memory score at pre-test (standard errors in parentheses) in each condition.

Table 3.1. The number of participants (the number of female participants in parentheses), the age range, and the mean age (standard deviations in parentheses), the verbal ability scores (standard errors in parentheses), the working memory score at pre-test (standard errors in parentheses) in each condition

<table>
<thead>
<tr>
<th>Condition</th>
<th>N</th>
<th>Age range</th>
<th>Mverb (SD)</th>
<th>MWMpre-test (SE)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Feedback with explanation</td>
<td>23 (15 female)</td>
<td>5;1 – 6;2</td>
<td>5.8 (0.28)</td>
<td>51.02 (0.10)</td>
</tr>
<tr>
<td>Feedback without explanation</td>
<td>23 (10 female)</td>
<td>5;2 – 6;8</td>
<td>5.8 (0.44)</td>
<td>53.70 (0.08)</td>
</tr>
<tr>
<td>No Feedback</td>
<td>26 (11 female)</td>
<td>5;2 – 6;8</td>
<td>5.4 (0.25)</td>
<td>53.01 (0.08)</td>
</tr>
<tr>
<td>Control</td>
<td>34 (19 female)</td>
<td>4;8 – 6;5</td>
<td>5.3 (0.35)</td>
<td>51.51 (0.13)</td>
</tr>
</tbody>
</table>

3.2.2. Design

Children were tested in three different experimental conditions and one control condition: (i) Feedback with explanation; (ii) Feedback without explanation; (iii) No feedback; (iv) Control. Each child was tested in five separate sessions, namely pre-test, training day 1, training day 2, post-test, and follow-up. There was at least one day intermission between the pre-test, training day 1, training day 2, and post-test sessions, and there was at least one week and at most nine days of intermission between the pre-test and the post-test sessions. The follow-up session was conducted four months after the pre-test session. Fig. 3.1 shows the design of the experiment.

At the pre-test, post-test and follow-up sessions, children were tested with a working memory task and three second-order false belief stories (1 ‘Three goals’, 1 ‘Decoy gift’, and 1 ‘Three locations’) in a random order. In the three different experimental conditions, children heard 6 second-order false belief stories (3 ‘Three goals’, 3 ‘Decoy Gift’) at each training session. After the third and the sixth second-order false belief stories, they heard one second-order true belief story (1 ‘Three goals’, 1 ‘Decoy gift’). In each second-order false belief story and...
Chapter 3: Five-year-old children's development of second-order false belief reasoning can be accelerated

3.2. Method

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 and further context of the stories. In the control condition, in both training sessions, children were tested with 7 neutral stories that do 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.

**Figure 3.1.** The design of our training study.

**Experimental conditions**
- Feedback with explanation
- Feedback without explanation
- No Feedback

**Tasks**
- 3 'Decoy gift' FB
- 3 'Three goals' FB
- 1 'Decoy gift' TB
- 1 'Three goals' TB

**Pre-test**
- Counting span task
- 3 Second-order FB
  1. 'Decoy gift' FB
  2. 'Three goals' FB
  3. 'Three locations' FB

**Training**

**Experimental conditions**
- Feedback with explanation
- Feedback without explanation
- No Feedback

**Tasks**
- 3 'Decoy gift' FB
- 3 'Three goals' FB
- 1 'Decoy gift' TB
- 1 'Three goals' TB

**Post-test**
- The same procedure as pre-test (different stories)

**Follow-up**
- The same procedure as pre-test (different stories)

**Session 1**
- 1-3 days

**Sessions 2 & 3**
- 1-3 days

**Session 4**
- 4 months

**Session 5**
- 4 months

**Figure 3.2.** The prototype 'Three locations' story, namely the 'Chocolate Bar' story (Illustration ©Avik Kumar Maitra)

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 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 pre-recorded control questions "Does Kevin know that Marieke put the chocolate into the toy box?", and "Does Marieke know that Kevin saw her put the chocolate into the toy box?" 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?" 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?" 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?" was asked together with the justification question "Why does she think that?"
Chapter 3: Five-year-old children’s development of second-order false belief reasoning can be accelerated

3.2. Method

Figure 3.3. The drawings of the prototype of ‘Three goals’ stories, namely ‘A Day Out’ story (Illustration ©Avik Kumar Maitra)

a) It is Robert’s birthday, so Robert’s dad promised to do something fun. Dad asks ‘Where do you want to go today?’ Robert says ‘The zoo’! Dad wants to call the zoo in order to make sure that it is open. He walks out of the room to get his phone.
b) Then, mother comes to the room. She asks Robert: “What are you doing today?” Robert says “We will go to the zoo!”. Mom says: “The zoo is not open today but you can also go to the swimming pool”. Robert thinks this is a good idea. He goes to find his dad to tell him that he wants to go to the swimming pool.

At this point, the control question “Does dad know that Robert wants to go to a swimming pool?” was asked.
c) Dad is alone in his room and he calls the zoo. He learns that the zoo is closed today. What now? He says to himself: “I know where to go, there is a very good movie in the cinema today, so I will call and book tickets for the movie”.

At this point the second control question “Does Robert know that his dad will go to a movie with him?” and the first-order false belief question “What does Robert think they are going to do today with his dad?” together with the justification question “Why does he think that?” were asked.
d) When dad has reserved the movie tickets, grandmother comes inside. She asks “What will you do with Robert today?”. Dad says: “We will go to the cinema”. Grandma says: “Oh, does Robert know what you are going to do today?”.

At this point, the control question (ignorance) “What does dad say to grandma?” was asked.

Subsequently, the last part of the story was told: “Then the grandma asks: ‘What does Robert think that you will do today?’”.

At this point, the second-order false belief question “What does Robert say to grandma?” together with the justification question “Why does he say that?” were asked.

Figure 3.4. The prototype example of ‘Decoy gift’ stories, namely the ‘Birthday Puppy’ story (Illustration ©Avik Kumar Maitra)

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” and the first-order false belief question “What does Rick think that his mom bought for him?” 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?” 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?” together with the justification question “Why does mum say that?” were asked.
3.2.3. Materials

3.2.3.1. Second-order false belief stories

We constructed 31 different second-order false belief stories of three different types: (i) 3 ‘Three locations’ stories, (ii) 14 ‘Three goals’ stories, (iii) 14 ‘Decoy gift’ stories. For all stories, children were asked a question that required second-order false belief attribution, as well as some control questions. In the literature, second-order false belief questions often have two possible answers, for example, two locations. We constructed ‘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 (i.e., zero-order, first-order, second-order).

Figure 3.2 shows the prototype example of ‘Three locations’ stories, namely the ‘Chocolate Bar’ story. 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 beliefs question “Where does Marieke think that Kevin will look for the chocolate?”: i) second-order ToM answer: “the drawer”, because Marieke thinks that Kevin thinks that the chocolate is in the drawer; ii) first-order ToM answer: “the toy box”, because Kevin thinks the chocolate is in the toy box; iii) zero-order ToM answer: “the TV stand”, because the chocolate is in the TV stand.

Figure 3.3 shows the prototype example of ‘Three goals’ stories, namely ‘A Day Out’. ‘Three goals’ stories included and extended the stories used in Hollebrandse, van Hout, and Hendriks’ (2014) study. Just like in the ‘Three location’ stories, there are three possible answers to the second-order false belief question: i) second-order ToM answer: “the zoo”, because Dad thinks that Robert thinks that they will go to the zoo; ii) first-order ToM answer: “the swimming pool”, because Robert thinks that they will go to the swimming pool; iii) zero-order ToM answer: “the cinema”, which is the real place to which they will go.

Figure 3.4 shows the prototype example of ‘Decoy gift’ stories, namely the ‘Birthday Puppy’ story. ‘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: i) second-order ToM answer: “a basketball”, because Mother thinks that Rick thinks that she bought a basketball; ii) zero-order ToM and first-order ToM answer: “a puppy”, because it is the real present and because Rick thinks that his mother bought a puppy.

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).

3.2.3.2. Second-order true belief stories

Second-order true belief stories were used only in the training sessions. In them, children were asked to answer a question that required attribution of a second-order true belief. Because we only trained children with ‘Decoy gift’ and ‘Three goals’ stories, we constructed the second-order true belief versions for only these types of stories, namely 2 ‘Decoy gift’ true belief stories and 2 ‘Three goals’ true belief stories. The true belief stories have the same structure as the false belief stories. 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 to the second-order false belief question in the corresponding false belief story (a basketball), because now the mother knows that the son knows that what she bought for him. 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.

3.2.3.3. Neutral stories

Neutral stories were presented to participants in the control condition in two training sessions (i.e., 7 stories in each training day). 14 neutral stories that have a similar length with the second-order false belief stories, and that do not involve theory of mind reasoning were selected from a children’s book called ‘Jip en Janneke’ by Annie-Schmidt, with Fiep 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 related to the episode of the story, in order to check if the children paid attention.
3.2.3.4. 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 at pre-test. We adapted the task from Towse et al.’s (1998) study. 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 number 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 number 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. In this scoring procedure: (i) the highest level (number of cards) for which at least two of the three trials were correct was noted as the main part of the score; (ii) if one of the three trials at the next level was correct, because this represents “half-way” toward the next span level, an additional 0.5 marks were given; (iii) the number of correct answers (correct item in the correct serial position) in each remaining trial was divided by the number of items to be recalled. The mean proportion of correct recalls for the incorrect trials was then derived. This value was multiplied by 0.5 and the product added to the total score obtained from procedures (i) and (ii).

3.2.3.5. 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 – 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.

3.2.4. Procedure

Children were tested individually in their school in a separate room by one of seven experimenters. 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-inch MacBook Pro OS X 10.10.5. All experimenters were trained before running the experiment in order to follow the same instructions. In the results, it turned out that there were no significant differences in scores between the children who were trained with different experimenters and adding experimenters as a random effect did not improve the linear mixed effect models. Therefore, we merged the data for the rest of the analysis. A child was almost always tested by the same experimenters at the pre-test, training day 1, training day 2, post-test 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 four months after the post-test. Thus, two of the remaining five experimenters tested these 17 of 106 children for the follow-up session. Each session took approximately 30 minutes. All of the sessions were recorded with QuickTime’s screen recording together with the 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 4 different second-order true belief stories, and (for the control condition) a pool of 14 different neutral stories. Drawings illustrating the story episodes of the stories were presented one by one, together with the corresponding audio recordings. The drawings remained visible when children were asked questions. As is usual in the previous studies, control questions were asked before the second-order false belief and second-order true belief questions 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, first-order false belief questions were asked before the second-order false belief questions, in order to make sure that the children did not have any major problems with first-order
false belief reasoning. 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 and the questions in the neutral stories that were used in the control condition, those questions were asked up to three times altogether by repeating the related story episodes. If a child still gave a wrong answer, it was coded as wrong. A child was never tested on the same story twice.

3.2.4.1. Pre-test, post-test 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. 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 pre-test, post-test, and follow-up sessions. Because children were not trained with ‘Three locations’ stories at the two training sessions, this type of stories was used to test whether children can generalize what they learned in the training sessions to another type of second-order false belief task.

3.2.4.2. Training Sessions

In the second and the third sessions (training day 1, training day 2), children in three of the experimental conditions were trained using six different second-order false belief stories (3 ‘Three goals’, and 3 ‘Decoy gift’) per training session. In addition to the second-order false belief stories in training sessions, children were tested with two second-order true belief stories in order to capture whether a child could have applied a simple strategy instead of reasoning about the second-order false belief questions. Each true belief story was presented after the 3 second-order false belief stories.

In the feedback with explanation condition, the feedback “Correct/Wrong” together with an explanation was provided in an interactive fashion. For example, the explanation was in the following form\(^\text{20}\) for the prototype of the ‘Decoy gift’ stories that was given in the Materials section: “Correct/Wrong, Rick saw the puppy in the basement but his mom didn’t see Rick, right? That’s why mom tells grandma that Rick thinks he will have a basketball. Because that is what the mom told Rick before, right?”. Similarly, the explanation was in the following form for the prototype of the ‘Three goals’ stories that was explained in the Materials section: “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?”

In the feedback without explanation condition, only the feedback “Correct/Wrong” was provided, together with the correct answer without any further explanation. In the no feedback condition and in the control condition, children did not get any feedback.

3.3. Results

Figure 3.5 shows (a) proportion of correct answers to the second-order false belief questions at pre-test, post-test and follow-up sessions and (b) the difference in the proportions between pre-test and post-test sessions for each condition. There is a considerable improvement of children’s scores from pre-test to post-test in the three experimental conditions: from 31% to 68% in the feedback with explanation condition; from 25% to 49% correct in the feedback without explanation condition; and from 33% to 55% in the no feedback condition, compared to a

\[^{20}\] Although 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 case of such small deviations.
Table 3.2. The percentage (and number in parentheses) of children showing an improvement, no change or deterioration in (a) answers to second-order false belief question and (b) answers to justification questions from pre-test to post-test sessions

<table>
<thead>
<tr>
<th></th>
<th>Improvement</th>
<th>Stability</th>
<th>Deterioration</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(a)</td>
<td>(b)</td>
<td>(a)</td>
</tr>
<tr>
<td>Feedback with explanation (N=23)</td>
<td>74 (17)</td>
<td>61 (14)</td>
<td>9 (2)</td>
</tr>
<tr>
<td>Feedback without explanation (N=23)</td>
<td>57 (13)</td>
<td>57 (13)</td>
<td>35 (8)</td>
</tr>
<tr>
<td>No feedback (N=26)</td>
<td>58 (16)</td>
<td>50 (13)</td>
<td>30 (8)</td>
</tr>
<tr>
<td>Control (N=34)</td>
<td>32 (11)</td>
<td>26 (9)</td>
<td>50 (17)</td>
</tr>
</tbody>
</table>

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 after 4 months from the pre-test session (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).

In this section, we first present the results of the training effects from pre-test to post-test in Subsection 3.3.1. Then we investigate the generalizability of the training effects by focusing on children’s improvements from pre-test to post-test sessions in Subsection 3.3.2. Subsequently, in Subsection 3.3.3, we present the stability of the training effects by focusing on the improvements from pre-test to follow-up sessions. Finally, in Subsection 3.3.4, we present the results of non-experimental questions.

3.3.1. Training effects from pre-test to post-test sessions

Following a similar pattern as Figure 3.5, Table 3.2 shows the percentage (and number in parentheses) of children showing an improvement, no change or deterioration in (a) answers to second-order false belief question and (b) answers to justification questions from pre-test to post-test sessions. As can be seen from Table 3.2, the most improvement in children’s answers to second-order false belief questions and justifications answers (together with less stability and less deterioration) occurs in the feedback with explanation group. In the feedback without explanation and no feedback conditions, children showed similar patterns of improvements. Moreover, as we expected, in the control condition, children’s improvement was much less, whilst their stability and deterioration were more compared to the children who were in one of the three experimental conditions.

Table 3.3. The estimates and z-values of the binomial mixed-effects model from pre-test to post-tests sessions

<table>
<thead>
<tr>
<th></th>
<th>( \beta )</th>
<th>SE</th>
<th>z</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 (Intercept)</td>
<td>-1.32</td>
<td>0.31</td>
<td>-4.24</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>2 Post-test</td>
<td>0.08</td>
<td>0.43</td>
<td>0.14</td>
<td>.86</td>
</tr>
<tr>
<td>3 Feedback without explanation</td>
<td>-0.67</td>
<td>0.46</td>
<td>-1.46</td>
<td>.14</td>
</tr>
<tr>
<td>4 Feedback with explanation</td>
<td>-0.35</td>
<td>0.42</td>
<td>-0.83</td>
<td>.41</td>
</tr>
<tr>
<td>5 No Feedback</td>
<td>0.18</td>
<td>0.37</td>
<td>0.48</td>
<td>.63</td>
</tr>
<tr>
<td>6 ‘Decoy gift’</td>
<td>1.63</td>
<td>0.25</td>
<td>6.51</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>7 ‘Three locations’</td>
<td>-0.22</td>
<td>0.35</td>
<td>-0.64</td>
<td>.52</td>
</tr>
<tr>
<td>8 Age</td>
<td>0.94</td>
<td>0.39</td>
<td>2.43</td>
<td>.02</td>
</tr>
<tr>
<td>9 Verbal ability</td>
<td>-0.01</td>
<td>0.02</td>
<td>-0.60</td>
<td>.55</td>
</tr>
<tr>
<td>10 Working memory</td>
<td>0.08</td>
<td>0.11</td>
<td>0.72</td>
<td>.47</td>
</tr>
<tr>
<td>11 Post-test x Feedback without explanation</td>
<td>1.51</td>
<td>0.70</td>
<td>2.17</td>
<td>.03</td>
</tr>
<tr>
<td>12 Post-test x Feedback with explanation</td>
<td>1.88</td>
<td>0.70</td>
<td>2.71</td>
<td>.007</td>
</tr>
<tr>
<td>13 Post-test x No Feedback</td>
<td>1.07</td>
<td>0.64</td>
<td>1.68</td>
<td>.09</td>
</tr>
<tr>
<td>14 Control x ‘Three locations’ x Post-test</td>
<td>0.34</td>
<td>0.63</td>
<td>0.54</td>
<td>.59</td>
</tr>
<tr>
<td>15 Feedback without explanation x ‘Three locations’ x Post-test</td>
<td>-0.86</td>
<td>0.78</td>
<td>-1.11</td>
<td>.27</td>
</tr>
<tr>
<td>16 Feedback with explanation x ‘Three locations’ x Post-test</td>
<td>0.28</td>
<td>0.72</td>
<td>0.39</td>
<td>.70</td>
</tr>
<tr>
<td>17 No Feedback x ‘Three locations’ x Post-test</td>
<td>-0.06</td>
<td>0.66</td>
<td>-0.09</td>
<td>.92</td>
</tr>
</tbody>
</table>

* The pre-test session, ‘Three goals’ stories, and control condition were used as base levels in the model (reference categories).

A binominal mixed effects model was fitted on the scores with the following effects: the main effects of and interaction between session (pre-test/post-test) 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 items was different from old 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 3.3 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 pre-test session, ‘Three goals’ stories, and control condition were used as base levels in the model (reference categories).

As can be seen from Table 3.3, children’s scores did not significantly improve from pre-test to post-test in the control condition (row 2). Children’s second-order
false belief scores significantly improved in the feedback with explanation condition and in the feedback without explanation condition compared to children's improvements in the control condition (rows 11 and 12). There was a marginally significant improvement from pre-test to post-test sessions in the no feedback condition compared to the control condition (row 13).

As expected, there was also a significant effect of age (row 8). We did not find a significant effect of children's verbal abilities (row 9) and working memory score at pre-test (row 10) on children's second-order false belief score. We interpret the results about the different story types in the following subsection about generalizability.

### 3.3.2. Generalizability of the training effect

In order to investigate the generalizability of the training effect, we focus on children's improvements from pre-test to post-test sessions in 'Three locations' stories. Note that, unlike 'Three goals' and 'Decoy gift' stories, we did not test children with 'Three locations' stories at the training sessions.

As can be seen from Table 3.3 (rows 6 and 7), while 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). 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 stories and children's 'Three locations' post-test scores in all of the conditions.

In Figure 3.6, we merged the stories that we used at the training sessions (‘Other Stories’), namely ‘Three goals’ and ‘Decoy gift’ stories, and compared with ‘Three locations’ stories. As can be seen from Figure 3.6, 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 pre-test to post-test (a rise of 29 percentage points in ‘Three locations’, and a rise of 27 percentage points in ‘Other Stories’ in the experimental conditions; compared to a rise of 8 percentage points in ‘Three locations’ and a rise of 4 percentage points in ‘Other Stories’ in the control condition). In more detail, Figure 3.7 shows children’s improvements in all of the story types of second-order false belief stories from pre-test to post-test sessions in all conditions. These results show that children were able to generalize what they learned at the training sessions to another story type that they did not train with, namely ‘Three locations’.

![Figure 3.6](#)

**Figure 3.6.** The comparison of children’s improvements in 'Three locations' vs. 'Other Stories' story types of second-order false belief stories from pre-test to post-test sessions.

![Figure 3.7](#)

**Figure 3.7.** Children's improvements in all of the story types of second-order false belief stories from pre-test to post-test sessions in all conditions. The green dashed horizontal line represents the chance level for the 'Decoy gift' stories (50%), and the blue and red dashed lines represent the chance level for the 'Three goals' and 'Three locations' stories (33%).
Table 3.4. The estimates and z-values of the binomial mixed-effects model for the stability of the training effect.

<table>
<thead>
<tr>
<th></th>
<th>B</th>
<th>SE</th>
<th>z</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>(Intercept)</td>
<td>-3.32</td>
<td>0.31</td>
<td>-10.67</td>
</tr>
<tr>
<td>2</td>
<td>Follow-up</td>
<td>0.35</td>
<td>0.39</td>
<td>0.92</td>
</tr>
<tr>
<td>3</td>
<td>Feedback without explanation</td>
<td>-0.71</td>
<td>0.45</td>
<td>-1.58</td>
</tr>
<tr>
<td>4</td>
<td>Feedback with explanation</td>
<td>-0.34</td>
<td>0.42</td>
<td>-0.81</td>
</tr>
<tr>
<td>5</td>
<td>No Feedback</td>
<td>0.16</td>
<td>0.38</td>
<td>0.41</td>
</tr>
<tr>
<td>6</td>
<td>'Decoy gift'</td>
<td>1.64</td>
<td>0.24</td>
<td>6.75</td>
</tr>
<tr>
<td>7</td>
<td>'Three locations'</td>
<td>-0.22</td>
<td>0.35</td>
<td>-0.64</td>
</tr>
<tr>
<td>8</td>
<td>Age</td>
<td>0.96</td>
<td>0.34</td>
<td>2.83</td>
</tr>
<tr>
<td>9</td>
<td>Verbal ability</td>
<td>0.002</td>
<td>0.02</td>
<td>0.10</td>
</tr>
<tr>
<td>10</td>
<td>Working memory</td>
<td>0.06</td>
<td>0.10</td>
<td>0.58</td>
</tr>
<tr>
<td>11</td>
<td>Follow-up x Feedback without explanation</td>
<td>0.76</td>
<td>0.62</td>
<td>1.22</td>
</tr>
<tr>
<td>12</td>
<td>Follow-up x Feedback with explanation</td>
<td>1.47</td>
<td>0.62</td>
<td>2.38</td>
</tr>
<tr>
<td>13</td>
<td>Follow-up x No Feedback</td>
<td>0.85</td>
<td>0.58</td>
<td>1.48</td>
</tr>
<tr>
<td>14</td>
<td>Control x 'Three locations' x Follow-up</td>
<td>1.38</td>
<td>0.57</td>
<td>2.41</td>
</tr>
<tr>
<td>15</td>
<td>Feedback without explanation x 'Three locations' x Follow-up</td>
<td>1.39</td>
<td>0.70</td>
<td>1.90</td>
</tr>
<tr>
<td>16</td>
<td>Feedback with explanation x 'Three locations' x Follow-up</td>
<td>0.92</td>
<td>0.71</td>
<td>1.30</td>
</tr>
<tr>
<td>17</td>
<td>No Feedback x 'Three locations' x Follow-up</td>
<td>1.62</td>
<td>0.68</td>
<td>2.40</td>
</tr>
</tbody>
</table>

3.3.3. Stability of the training effects: Improvements from pre-test to follow-up sessions

As can be seen from Figure 3.5, for all the conditions, children’s scores on second-order false belief stories improved from the pre-test to a follow-up session, which was 4 months after the pre-test session (a rise of 42 percentage points in the feedback with explanation condition; a rise of 31 percentage points in the feedback without explanation condition; a rise of 35 percentage points in the no feedback condition; compared to a rise of 17 percentage points in the control condition).

Similar to the fitted binomial linear mixed effect model that we presented in Table 3.3 (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 (pre-test/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 pre-test 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. Table 3.4 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 3.4, in control condition, children’s second-order false belief scores did not significantly improve from pre-test to follow-up sessions (row 2). There was a significant difference from pre-test to follow-up sessions between the control condition and the feedback with explanation (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 (row 11, 12).

Moreover, children’s improvements from pre-test to follow-up sessions in ‘Three locations’ stories were more than children’s improvements in the trained stories in all the conditions (rows 15, 16, 17).

Finally, similar to the findings from pre-test to post-test sessions that is shown in Table 3.3, while age had a significant effect, children’s verbal abilities and working memory score at pre-test did not have a significant effect on children’s second-order false belief score from pre-test to follow-up sessions.

3.3.4. Non-experimental questions

Table 3.5 shows the percentages of correct answers (standard errors in parentheses) for the non-experimental questions.

<table>
<thead>
<tr>
<th>Question type</th>
<th>Total number of questions</th>
<th>Correct answers</th>
</tr>
</thead>
<tbody>
<tr>
<td>Control questions</td>
<td>4968</td>
<td>98% (0.002)</td>
</tr>
<tr>
<td>First-order false belief</td>
<td>1811</td>
<td>95% (0.005)</td>
</tr>
<tr>
<td>First-order true belief</td>
<td>286</td>
<td>97% (0.01)</td>
</tr>
<tr>
<td>Second-order true belief</td>
<td>286</td>
<td>92% (0.02)</td>
</tr>
<tr>
<td>Questions in neutral stories in the control condition</td>
<td>1900</td>
<td>95% (0.005)</td>
</tr>
</tbody>
</table>
questions as well as attributing first-order false beliefs. Moreover, in order to make sure that children did not use a simple strategy instead of reasoning about the second-order false belief questions, we investigated children’s performance on second-order true belief questions.

As can be seen from Table 3.5, 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 second-order true belief answers shows that children did not use simple strategy instead of applying second-order ToM reasoning.

3.4. Discussion and future research

For the first time in the literature, the roles of different types of feedback, namely feedback with explanations, feedback without explanation, and no feedback have been studied, to accelerate 5- to 6-year-olds’ development of second-order false belief reasoning. Crucially, the design of our study covers and extends the important suggestions of Hoffman et al.’s (2016) meta-analysis of the training studies of ToM, by: 1) controlling for age, working memory and verbal abilities; 2) testing children in a follow-up session after 4 months from the post-test session; 3) controlling for children’s simple strategy use instead of reasoning about others’ minds by testing them with second-order true belief stories; 4) testing children with a ToM task that is not used in the training sessions in order to see whether children can generalize what they have learned; and 5) using an active control condition.

Overall, our results draw attention to the similarities and the differences in first-order and second-order ToM development. We first discuss the training effect and its stability. Subsequently, we discuss the results about the generalizability of the training effect.

3.4.1. The training effects and their stability

As we predicted, children in the feedback with explanation condition made greater gains in second-order ToM from pre-test to post-test 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 ToM training studies (Appleton & Reddy, 1996; Lecce et al., 2014a; Clements et al., 2000; Melot & Angeard, 2003) and with the second-order ToM studies that tested 9- to 10-year-olds with more advanced ToM tasks (Bianco et al., 2015; Lecce et al., 2014b). This result of 5- to 6-year olds shows the validity of our training study and fills the gap in the ToM literature between the preschool children and middle childhood.

Our second prediction that, there would also be a significant improvement in the feedback without explanation condition from pre-test to post-test sessions compared to control condition, was also confirmed. In this condition, the feedback “Correct/Wrong” together with the correct answer was provided without providing further explanation. This result is important given the previous first-order ToM training studies that showed that 3- to 4-year-olds did not improve from pre-test to post-test session if they received “Correct/Wrong” feedback without explanation (Clements et al., 2000). An explanation for the improvement that we found derives from a previous computational cognitive modeling study. Arslan et al. (2013; 2017a) predicted that once children are able to attribute first-order false beliefs to an agent, they initially use a first-order reasoning strategy in second-order false belief tasks. The model predicts that later on, with repeated exposure to second-order ToM reasoning and with the help of feedback “Correct/Wrong”, children can revise their first-order ToM strategy to one level higher, namely to a second-order ToM strategy. Note that Clements et al. (2000) did not provide the correct answer after the feedback “Correct/Wrong”. Further research is needed with a training study of first-order false belief reasoning to test whether 3-year-old children’s performance can be improved from pre-test to post-test when the correct answer is provided together with the feedback “Correct/Wrong” without any further explanations.

Our training study also shows other interesting and somewhat unexpected results. Children’s performance from pre-test to post-test sessions also improved in the no feedback condition, and there was a marginally significant difference between the no feedback and control conditions ($B = 1.07, SE = 0.64, p = .09$). Considering Melot and Angeard’s (2003) and Clements et al.’s (2000) findings that training 4-year-old children with first-order false belief tasks without providing any feedback and providing feedback without explanation did not improve children’s performance, our results highlight the difference between first-order ToM and second-order ToM development. For the unexpected improvement in the no feedback condition, we surmise that exposing children to second-order ToM and asking them second-order false belief questions, together with the justification questions “Why?” helps children to reflect about their own judgments. Thus, asking justification questions helps children to revise their wrong strategy to a correct second-order ToM strategy. This argument needs to be tested with another training study in which children are trained on second-order ToM with no feedback, however, this time without asking the justification questions.
As can be seen from Figure 3.5 and Table 3.4, for all the conditions, children’s scores on second-order false belief stories continued to improve 4 months after the pre-test session. Importantly, 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 at the 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. The small improvement from pre-test 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 9 times (3 stories each at pre-test, post-test, and follow-up sessions).

Moreover, we did not find any effect of children’s verbal abilities and working memory score at pre-test on children’s second-order false belief score. The insignificant effect of working memory is in line with Hasselhorn et al.’s (2005) study. We interpret the insignificant 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 just to 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 needs more complex working memory strategies might predict children’s second-order false belief scores. 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 ToM development. Further research is needed to verify this prediction.

3.4.2. Generalizability of the training effects

Before commenting on the generalizability of the training effects, it is worth discussing the differences between the three story types that we used in our study, namely ‘Decoy gift’, ‘Three goals’, and ‘Three locations’. As we discussed in Subsection 3.2.3.1, ‘Decoy gift’ stories are less complex than the ‘Three goals’ and ‘Three locations’ stories because ‘Decoy gift’ stories have two possible answers compared to three possible answers 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 pre-test session.

Note that we did not use ‘Three locations’ stories at the training sessions in order to test the generalizability of the training effect. In the experimental conditions, children’s improvements in the stories that children trained on (i.e., ‘Three goals’ and ‘Decoy gift’ stories) did not significantly differ from their improvements in ‘Three locations’ stories from pre-test to post-test sessions (see Figure 3.6 and Table 3.3). These results together show the generalizability of the training effect.

Moreover, as can be seen in Table 3.4, we found an interesting and unexpected result about children’s improvements in ‘Three locations’ stories from pre-test to follow-up sessions. In all conditions, children’s improvements from pre-test 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 of 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 ‘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 (Hollebrandse & Roepen, 2014; de Villiers, Hobbs, & Hollebrandse, 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.

In addition to the above-mentioned strengths of the design and the novelty of our findings, our study also has a number of limitations that should be acknowledged. First, the number of children in each condition is not large. Second, the variability in the socioeconomic status of the children was limited. Third, although
we used another second-order false belief story type to assess the generalizability of our training program, further research is needed 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.

3.5. Conclusions

Our results showed that there are considerable improvements in children’s scores from pre-test to post-test in the three experimental conditions: from 31% to 68% in the feedback with explanation condition; from 25% to 49% in the feedback without explanation condition; and from 33% to 55% in the no feedback condition, compared to a small improvement in the control condition (from 29% to 35%). The considerable improvements in the experimental conditions are not due to children’s age, working memory, verbal abilities, or using a simple strategy instead of attributing second-order false beliefs.

Our findings that 5-year-olds’ second-order false belief reasoning can be accelerated with different kinds of feedback both corroborate and contrast with the existing first-order ToM training studies. The improvements in the feedback with explanation condition are in line with the first-order ToM training studies that tested 3- to 5-year-olds and the advanced ToM studies that tested 9- to 10-year-olds. On the other hand, children’s second-order false belief improvement in the feedback without explanation and in the no feedback conditions are contrary to the first-order ToM literature that showed that 3-year-olds’ first-order ToM development could not be accelerated by training with feedback without explanations or without any feedback. Importantly, children can generalize the training effect to a story type on which they did not train and the training effect is stable over 4 months.

Based on our results, we suggest that children can be helped over 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 explanations about their wrong answers.