# Teaching Children to Attribute Second-order False Beliefs: A Training Study with Feedback

**Burcu Arslan[1] (barslan.cogs@gmail.com), Rineke Verbrugge[1] (l.c.verbrugge@rug.nl), Niels Taatgen[1] (n.a.taatgen@rug.nl), Bart Hollebrandse[2] (b.hollebrandse@rug.nl)**

[1] Institute of Artificial Intelligence, University of Groningen, P.O. Box 407, 9700 AK Groningen, The Netherlands

[2]Faculty of Arts, University of Groningen, P.O. Box 716, 9700 AS Groningen, The Netherlands

**Abstract**

The ability to reason about another person's mental states, such as belief, desires and knowledge – first-order theory of mind – develops between the ages three and four. On the other hand, children need one or two more years to reason about a person who reasons about another person – second-order theory of mind. Is it possible to accelerate the development of theory of mind? There are several training studies that showed that it is possible to teach preschool children to pass first-order false belief tasks. However, the literature is missing analogous training effects for school-age children with respect to second-order false belief tasks. In this study, we focus on the role of feedback in the development of second-order false belief reasoning in two different conditions in children between the ages five and six: (i) feedback with explanation, (ii) feedback without explanation. Children's performance improved in both conditions. Previous theories suggest either that children's development of second-order theory of mind requires conceptual changes or that 4-5 year old children have cognitive constraints that need to be overcome in order for them to be able to apply second-order theory of mind. In line with our findings, however, we argue that five-year-old children who cannot yet pass the second-order false belief task reason about the false belief questions based on the reasoning strategy that they most frequently use in daily life (i.e. first-order or zero-order theory of mind). Moreover, we argue that most of the time children can revise their wrong reasoning strategy and change to the correct second-order reasoning strategy based on repeated exposure to the feedback "Correct/Wrong" together with the correct answer.

**Keywords:** Second-order theory of mind, false belief reasoning, feedback, training.

## Introduction

To understand and predict the behavior of others, people regularly reason about other people's mental states, such as beliefs, desires, knowledge and intentions. This ability is called theory of mind (ToM) (Premack & Woodruff, 1978). If we reason about world facts such as the location of an object, we do not attribute any mental states to another person (zero-order reasoning). However, if we reason about what a colleague believes about the location of the object, we are attributing a belief to that colleague (first-order ToM). In more complex social situations, we do not only reason about what a colleague believes but also we reason about what a colleague believes that we think (second-order ToM) and so forth. Those different orders of reasoning develop with age.

The most studied task for assessing the development of ToM is called the *false belief task* (Wimmer & Perner, 1983). In the verbal first-order false belief task, a child is expected to answer a question about a protagonist who has a false belief about a situation, while the child itself has a true belief about the same situation. For the second-order false belief task, children are expected to answer a question about what a protagonist thinks about another protagonist's beliefs or knowledge, such as "Where does Ayla think that Murat will look for the chocolate?" (see Materials section for more details about second-order false belief stories and questions). After children can pass first-order false belief tasks around the age of 4 (Wellman, Cross & Watson, 2001), it takes them one or two further years to pass the second-order false belief task (Perner & Wimmer, 1985; Sullivan et al., 1994).

Is it possible to accelerate the development of ToM? There are several training studies showing that it is possible to teach pre-school children to pass first-order false belief tasks (see Kloo & Perner, 2008 for a review). The general procedure in those successful first-order ToM training studies starts by pre-testing children who are on the verge of developing first-order ToM to make sure that they have not developed it yet. Subsequently, the children are trained with false belief tasks, either with or without feedback (Clements, Rustin, & McCallum, 2000; Melot & Angeard, 2003) to investigate the role of feedback. Alternatively, in order to examine the contributing factors in ToM development, children are exposed to tasks testing different cognitive abilities, such as language (Hale & Tager-Flusberg, 2003) and executive functions, such as inhibition and working memory (Kloo & Perner, 2003). The studies that tested the role of feedback showed that children's performance increased when they had been trained with detailed explanations, but not when they had only received the feedback "Correct/Wrong" without any explanation. These results are in line with theories proposing that children's development of first-order theory of mind depends on conceptual change (Gopnik & Wellman, 2012).

On the other hand, the literature is still missing analogous training studies that examine the effect of feedback, language, and executive functions on the development of second-order theory of mind during the primary school years, except for a recent training study with 9- and 10-year-olds that highlighted the important role of conversation

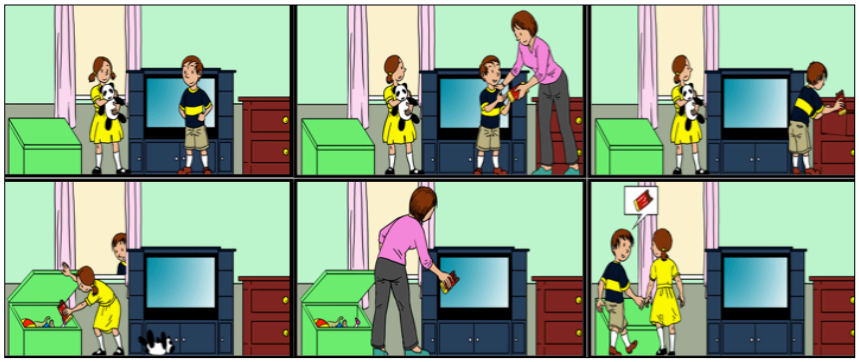

Figure 1: The Chocolate Bar story (see Materials section; Illustrator ©Avik Kumar Maitra)

about mental states in a nonliteral interpretation ToM task (Lecce et al., 2014). Before the empirical training study that we report here, we performed a computational cognitive modeling study, which predicted that children who are between the ages of 5 and 6 can learn to pass second-order false belief stories with the help of only the feedback "Correct/Wrong", without explanation (Arslan, Taatgen, & Verbrugge, 2013). The cognitive model starts to reason about a second-order false belief question (e.g. "Where does Ayla think that Murat will look for the chocolate?") from its own perspective (zero-order reasoning) and gives an answer accordingly (i.e. the real location of the chocolate).

The theoretical explanation behind our modeling choice is that young children experience reasoning about world facts predominantly from their own perspective. However, after the model repeatedly receives the feedback "Wrong", the model increments its strategy one level up (first-order reasoning), and gives an answer accordingly (i.e. what Murat thinks about the location of the chocolate). Because this is still not the correct answer, the model still gets the feedback "Wrong". It then again increments its strategy by one level (second-order reasoning), and, finally, it gives the correct answer (i.e. where Ayla thinks that Murat will look for the chocolate). This time, because the answer is correct, the model gets the feedback "Correct", and it stabilizes its second-order strategy.

Thus, our model suggests that in principle, children around the age of 5 can pass second-order false belief tasks. However, the problem that they still encounter is that, while their conceptual development has advanced far enough to *understand* that people can have different second-level perspectives, they are not *used* to reasoning about second-order mental states in their daily lives yet. Thus, we propose that by getting sufficient experience and by getting the feedback "Correct/Wrong", children can revise their reasoning strategy and can pass second-order false belief tasks.

Therefore, the goal of the current study is to test our model's prediction (Arslan, Taatgen, and Verbrugge, 2013) that children between the ages 5 and 6 can learn to pass second-order false belief tasks with the help of feedback.

## Method

### Participants

A sample of 51 Dutch 5 to 6 year-old children from predominantly upper-middle-class families was recruited from a primary school in Groningen, the Netherlands, and tested individually in their school in a separate room. The children were pre-tested to ensure that they had not yet fully developed second-order reasoning about false beliefs. Accordingly, four children who gave correct answers for all of the three second-order false belief questions contained in pre-test were excluded from our analysis, as well as one child who experienced technical problems in one run of the experiment. Therefore, the analysis included the results of 23 children in the 'feedback with explanation' group (15 female, $M_{age}$=5.8 years, SE=0.06, range: 5.1 – 6.2), and 23 children in the 'feedback without explanation' group (10 female, $M_{age}$=5.8 years, SE=0.09, range: 5.2 – 6.8).

### Materials

**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 1 shows the prototype example of a 'three-locations' story, namely the Chocolate Bar story. In each story 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.

*'Three locations'* stories were constructed based on Flobbe and colleagues' (2008) Chocolate Bar story (see Figure 1), as follows. Two siblings play in a room. The mother gives a chocolate bar to her son Murat but not to her

daughter Ayla and then leaves the room. Murat eats some of the chocolate, puts the remainder into the drawer, and leaves the room as well. Because he did not give any chocolate to his sister, Ayla wants to play a trick on him. She takes the chocolate from the drawer and puts it into the toy box. While she is hiding the chocolate in the toy box, Murat is passing by the window and sees Ayla put the chocolate into the toy box; however, Ayla does not see Murat. After that, Ayla leaves the room, too. Then the mother enters to tidy up the room; she finds the chocolate in the toy box, and she places it on the TV stand. The experimenter asks the participant the second-order false belief question: "Where does Ayla think that her brother Murat will look for the chocolate?". There are three possible locations to be reported to this question: the drawer (second-order answer), the toy box (first-order answer), and the TV stand (zero-order answer).

*'Three goals'* stories included and extended the stories used in Hollebrandse, van Hout, and Hendriks' (2014) study. One of the examples of this story type is as follows: Ruben and Myrthe play in their room. Myrthe tells Ruben that she will go to buy chocolate-chip cookies from the bake sale at the church and she leaves the house. After that, their mother comes home and tells Ruben that she just visited the bake sale. Ruben asks his mother whether they have chocolate-chip cookies at the bake sale. The mother says, "No, they have only apple pies". Then Ruben says, "Oh, then Myrthe will buy an apple pie", Meanwhile, Myrthe is at the bake sale and asks for the chocolate-chip cookies. The saleswoman says, "Sorry, we only have muffins". Myrthe buys some muffins and goes back home. While she is on her way home, she meets the mailman and tells him that she bought some muffins for her brother Ruben. The mailman asks her what Ruben thinks that she bought. At this point, the experimenter asks the participant "What was Myrthe's answer to the mailman?" There are three possible answers that children might report: chocolate-chip cookies, which Myrthe told Ruben initially (second-order answer); an apple pie, which the mother told Ruben (first-order answer); and muffins, which Myrthe really bought (zero-order answer)**.**

*'Decoy gift'* stories were constructed based on Sullivan and colleagues' (1994) Birthday Puppy story, on the following lines: A father deliberately lies to his daughter about the present he will give her for her birthday, in order to surprise her later. However, when the father is not in the room, the daughter finds her real birthday present. In the meantime, the daughter's grandmother calls the father and asks him what the daughter thinks that she is getting for her birthday. In this story, there are two possible objects that the participants might report: the decoy gift about which the parent deliberately lied to his child (second-order answer), and the real present that the child found (zero-order and first-order answer).

**Second-order true belief stories.** In the true belief stories, children were asked to answer a question that required attribution of a second-order true belief. We constructed two 'decoy gift' stories and two 'three goals' 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 daughter finds her real birthday present, but the father is also in the room and they jointly attend the present. Therefore, this time the correct answer to the second-order true belief question is not the same as the second-order false belief answer.

**Counting span task.** This task is a simple working memory task. We adapted it from Towse and colleagues' (1998) study. In the task, there are red triangles and blue squares on each card. 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. After being sure that children understood the instructions, the real experiment started.

In the first level, after two cards, the children were asked to report the total number of blue shapes 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 was increased by one. For the scoring, we adapted the same criteria of Towse and colleagues' (1998) study. In this scoring procedure, the highest level (number of cards) for which two of the three trials were correct was noted as the main part of the score. Moreover, the number of a child's correct answers in the next level was included as the secondary part of its score.

## Procedure

All the stimuli were presented to the children in a 15-inch MacBook Pro and were implemented with Psychopy2 v.1.78.01. Each child was tested on four different days, which are referred to as sessions for the rest of this paper. There was at least one day and at most three days of intermission between the sessions, and there was at least one week of intermission between the first and the fourth sessions. Each session took approximately 30 minutes.

In the first session (pre-test), children were pre-tested in order to test that they did not pass all of the three second-order false belief stories (1 'three goals' 1 'decoy gift', and 1 'three locations'). If a child gave correct answers for all of them, their score was coded as 3, and they were excluded from the data analysis. In addition to the stories, children were tested with the counting span task. The presentation of the order of the tasks was randomized.

In the second and the third sessions (training sessions), children were trained using six different second-order

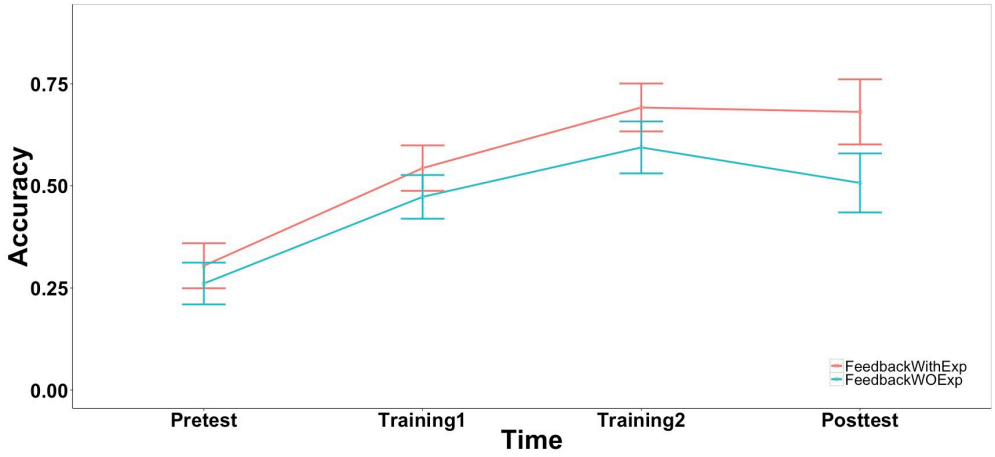


Figure 3: Children's improvement in second-order false belief scores from pre- to post-tests (error bars represent SEs).

false belief stories (3 'three goals', and 3 'decoy gift') in each training session. Therefore, the maximum score for each training session was 6. In addition to the second-order false belief stories, the 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 questions. The true belief stories were always asked after the 3 false belief stories. Because children were not trained with 'three locations' stories during the two training sessions, this type of story was used to test the transfer effect of the training, from pre-test to post-test. Finally, in the fourth session (post-test), children were post-tested using exactly the same procedure as used in the pre-test[1].

Stories were pseudo-randomly drawn from a pool that contained 31 different false belief stories and a pool of 4 different true belief stories. Drawings illustrating the story episodes were presented one by one, together with the corresponding audio recordings. The drawings remained visible throughout the story. Control questions were asked before the second-order belief questions, to test that children did not have major memory and linguistic problems about the stories and the structure of the questions. A child was never tested on the same story twice.

Children were tested in two different experimental conditions: (i) feedback with explanation; (ii) feedback without explanation. For the *feedback with explanation* group, the feedback "Correct/Wrong" together with an explanation was provided in an interactive fashion (e.g. "Correct/Wrong, Did Murat see that Ayla put the chocolate into the toy box? Yes, right? Did Ayla see Murat? No, right? That is why Ayla thinks that Murat will look for the chocolate in the place where he put it, and that is the drawer, isn't it?"). For the *feedback without explanation* group, only the feedback "Correct/Wrong" was provided, together with the correct answer without any further explanation. No feedback was provided to any child during pre- and post-tests.

[1] Children were also tested with a theory of mind game during pre- and post-test in order to investigate the far transfer effect. In this game, children were expected to reason about the computer's decision (first-order ToM) and about the computer's belief about their own decision (second-order ToM). However, the task was too hard for the 5-6 years olds. For this reason, we do not present the game and its results here.

## Results

We used binomial linear mixed effect models by using the glmer function in the lme4 package (Bates, Maechler, Bolker & Walker, 2014) for the statistical software R. The estimates of the coefficients are reported in log odds.

### Second-order false belief stories

As can be seen from Figure 3, there is a considerable improvement of children's scores from pre-test to post-test (from 31% to 69% correct in the feedback with explanation condition, and from 23% to 51% correct in the feedback without explanation condition). Figure 4 shows children's improvements in second-order false belief scores from pre- to post-tests for different types of stories in both experimental conditions.

A binominal mixed effects model was fitted on the scores with an interaction between test condition (pre-test/post-test), experimental condition (feedback with and without explanation), and story type ('decoy gift', 'three goals', and 'three locations'). As random effects, we had intercepts for subjects, and random slopes for time per subject correlated with the random intercepts. Table 1 lists the estimates and z-statistics of the mixed-effects model.

There are significant main effects for post-test and 'decoy gift' stories. In the feedback with explanation condition, children's performance increased equally for all three story types. In contrast, as can be seen from Figure 4, in the feedback without explanation condition children's scores of 'three locations' stories did not improve as much as for the other two story types, which were used during training sessions. However, we couldn't find any statistically significant evidence to show this effect.

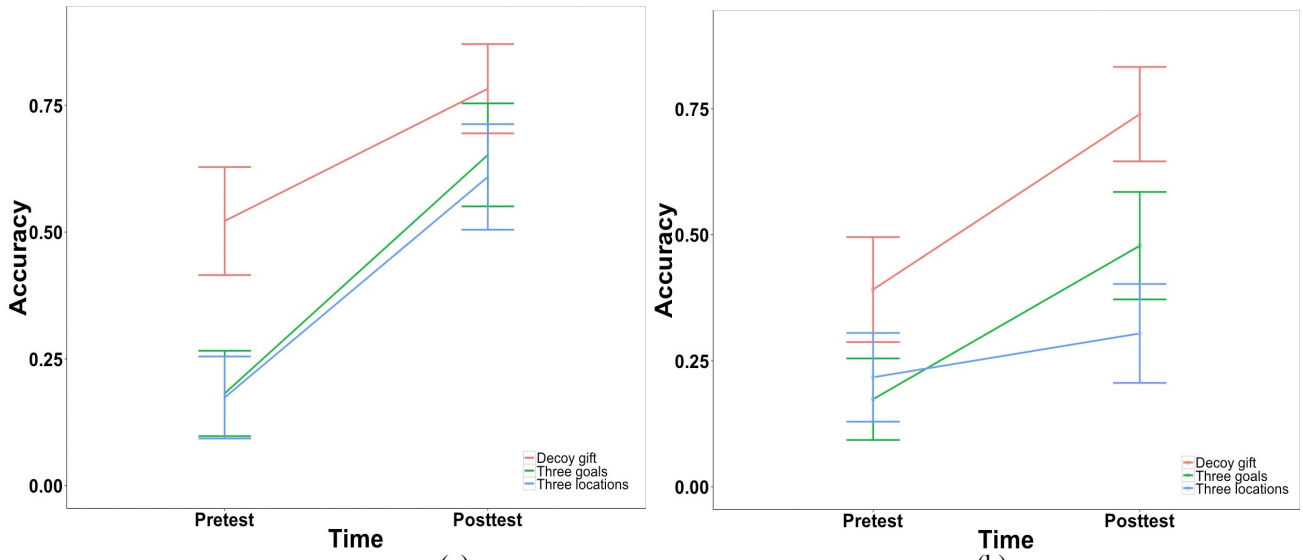


Figure 4: Children's improvements in second-order false belief scores from pre- to post-tests for different types of stories in (a) the feedback with explanation group and (b) the feedback without explanation group.

Table 1: The estimates and z-values of the mixed-effects model for pre- and post-tests

| | *β* | SE | *z* | *p* |
|---|---|---|---|---|
| Intercept | -1.56 | 0.55 | -2.83 | .005** |
| Post-test | 2.32 | 0.90 | 2.57 | .010* |
| Without explanation | 0.28 | 0.75 | 0.37 | .71 |
| 'Decoy gift' | 1.65 | 0.69 | 2.39 | .017* |
| 'Three goals' | 0.05 | 0.78 | 0.06 | .949 |
| Posttest: without explanation | -2.41 | 1.29 | -1.87 | .062· |
| Posttest: 'decoy gift' | -0.24 | 1.13 | -0.21 | .832 |
| Posttest: 'three goals' | 0.27 | 1.12 | 0.24 | .810 |
| Without explanation: 'decoy gift' | -0.81 | 0.96 | -0.84 | .398 |
| Without explanation: 'three goals' | -0.33 | 1.08 | -0.30 | .762 |
| Posttest: without explanation: 'decoy gift' | 2.50 | 1.58 | 1.59 | .112 |
| Posttest: without explanation: 'three goals' | 1.24 | 1.57 | 0.79 | .430 |

### Second-order true belief stories

In order to make sure that children did not use a simple strategy instead of reasoning about the questions, we investigated children's performance on second-order true belief questions. Overall, the true belief questions were answered correctly in 85% of cases in both conditions. The high proportions of correct answers suggest that children did not use a simple strategy (such as "take the object in the top-left picture") in the false belief tasks, otherwise they would probably have used the same (then incorrect) answers for the true-belief questions.

### Counting span task

To see whether counting span task scores predict the false belief scores and the learning effect, we added the counting span score with its interaction with time to the binomial linear mixed effect model. We couldn't find any significant effect of counting span task scores on the second-order false belief scores.

## General Discussion, Conclusions and Future Directions

To the best our knowledge, we have shown for the first time in the literature that children's performance on the second-order false belief task can be improved with the help of both feedback with explanation and feedback without explanation. Moreover, our finding that children performed around 85% correct on the true belief stories suggests that the training effect cannot be interpreted simply by assuming that children were applying a simple strategy instead of learning to attribute second-order false beliefs.

Because we provided detailed explanations with interactive feedback in the *feedback with explanation* group, the children's improvement in that group was expected, also considering Clements, Rustin, and McCallum's (2000) and Melot and Angeard's (2003) studies that found a positive effect of training with feedback with explanations in first-order theory of mind tasks.

Moreover, our results about the improvement in the *feedback without explanation* group are in line with our previous computational cognitive model's predictions that children's performance will improve with the help of feedback without explanation, simply on the basis of "Correct"/"Wrong" feedback (Arslan, Taatgen & Verbrugge, 2013). On the other hand, this result differs from Clements, Rustin, and McCallum's (2000) finding in the

first-order false belief reasoning domain, namely that explanation is necessary for training effects. Note that Clements and colleagues (2000) did not provide the correct answer to the children after giving the feedback ("Correct/Wrong"), in contrast to our study.

What does it mean to have a training effect in both the feedback with and without explanation conditions? The improvements in the feedback with explanation condition suggest that 5- to 6-year olds who cannot yet pass the second-order false belief tasks before the experiment are actually able to pass those tasks with the help of related explanations, and there is no cognitive constraint to prevent them passing the second-order false belief tasks. In addition, the improvements in the feedback without explanation condition suggest that even if children do not receive any explanation, they can still revise their strategy and can pass the second-order false belief tasks with the help of feedback "Correct/Wrong" together with the correct answer. This result, together with our previous computational cognitive model, can be interpreted as follows: children might be able to make the necessary reasoning steps, however, they might not be used to applying those strategies in practice.

We are currently running the third control condition of the experiment, in which we are training children with second-order false belief tasks without any feedback. Because we do not have the data for this condition in this paper, we cannot rule out another possible interpretation of the training effect. That is, just hearing second-order false belief stories and answering the related questions might help children. Moreover, in order to be able to conclude that the training effect is not just for a short time period, we are also conducting a follow-up test in which children are tested again a couple of months later than the actual training sessions.

## Acknowledgments

We are grateful to the Netherlands Organization for Scientific Research for Vici grant NWO-277-80-01, awarded to Rineke Verbrugge. We are also thankful to the managers and teachers of Joseph Haydn School in Groningen and to the children's families who allowed us to carry on this study. Finally, we would like to thank Avik Kumar Maitra for the illustrations of the stories, Bea Valkenier for being the voice of the stories, Maximilian Seidler for running the pilot study and for the code of the experiment, and Harmen de Weerd for his continuous support.

## References

Arslan, B., Taatgen, N. A. & Verbrugge, R. (2013). Modeling developmental transitions in reasoning about false beliefs of others. In R. West & T. Stewart (eds.), *Proceedings of the 12th International Conference on Cognitive Modeling*, Ottawa: Carleton University. 77-82.

Arslan, B., Hohenberger, A., & Verbrugge, R. (*Submitted*). The role of language and memory in the development of second-order theory of mind.

Bates D., Maechler M., Bolker B.M. & Walker S. (2014). lme4: Linear mixed-effects models using Eigen and S4. ArXiv e-print; submitted to *Journal of Statistical Software*, http://arxiv.org/abs/1406.5823.

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.

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 (4), 417-442.

Gopnik, A., Wellman, H.M., 2012. Reconstructing constructivism: Causal models, Bayesian learning mechanisms, and the theory theory. *Psychological Bulletin* 138, 1085–1108.

Hale, C. M., & Tager-Flusberg, H. (2003). The influence of language on theory of mind: A training study. *Developmental Science*, 6, 346–359.

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 (3), 321-333.

Kloo, D., & Perner, J. (2003). Training transfer between card sorting and false belief understanding: Helping children apply conflicting descriptions. *Child Development,* 74 (6), 1823-1839.

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.

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.

Melot, A. N., & Angeard, N. (2003). Theory of mind: Is training contagious? *Developmental Science*, 6, 178–184.

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.

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

Sullivan, K., Zaitchik, D. & Tager-Flusberg. H. (1994). Preschoolers can attribute second-order beliefs. *Developmental Psychology*, 30 (3). 395-402.

Towse, J. N., Hitch, G. J., & Hutton, U. (1998). A reevaluation of working memory capacity in children. *Journal of Memory and Language,* 39 (2), 195-217.

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.

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.