Reasoning about diamonds, gravity and mental states: The cognitive costs of theory of mind

Ben Meijering[1] (b.meijering@rug.nl), Hedderik van Rijn[2], Niels A. Taatgen[1], and Rineke Verbrugge[1]

[1]Institute of Artificial Intelligence, PO Box 407
9700 AK Groningen, The Netherlands
[2]Department of Psychology, Grote Kruisstraat 2/1
NL-9712 TS Groningen

Abstract

Theory of mind (ToM) is required when reasoning about mental states such as knowledge, beliefs, desires, and intentions. Many complex reasoning tasks require domain-general cognitive resources such as planning, resistance to interference, and working memory. In this paper we present a study of the additional cognitive costs of reasoning about mental states. We presented participants with sequential games in which they have to reason about another player. In the so-called player condition, the other player is reasoning about the participant, whereas in the so-called balance condition, the other player is reasoning about a balance scale. Both types of games require the same comparisons, but only differ in the required depth of ToM reasoning. Games in the player condition require one additional switch between perspectives. The results show that participants make different types of mistakes in the player condition as compared to the balance condition. This finding implies a different reasoning process when reasoning about mental states. The results also show faster decreasing reaction times in the balance condition than in the player condition. Based on these findings, we argue that reasoning about mental states requires unique cognitive resources.

Keywords: Theory of mind; perspective taking; decision making; sequential games; social cognition.

Introduction

In many social interactions we reason about one another. If, for example, our decisions or outcomes depend on someone else's actions, we try to predict what the other will do. Predicting the other's actions requires an understanding of how behaviors are caused by mental states such as beliefs, desires, goals, et cetera. Such an understanding is often referred to as *theory of mind* (Premack & Woodruff, 1978).

A theory of mind, or ToM, is starting to develop around the age of three to four years (Wellman, Cross, & Watson, 2001; Wimmer & Perner, 1983). However, younger infants already are susceptible to others' mental states (Onishi & Baillargeon, 2005). One possible explanation is that they are able to read others' behavior, but cannot yet explicitly reason about the underlying mental states. Only after many interactions, reading many distinct behaviors, do children start to develop a theory of how behaviors generally correspond with beliefs, desires, intentions, et cetera.

So far, we have introduced ToM as being a *theory* (Gopnik & Wellman, 1992). However, we do not want to exclude another definition of ToM that considers it to be an *ability* or skill to reason about mental states of oneself and others (Apperly, 2011; Leslie, Friedman, & German, 2004; Van Rij, Van Rijn, & Hendriks, 2010; Wimmer & Perner, 1983). In fact, a theory alone would not suffice when reasoning about others' mental states. Such reasoning is an entire process of generating many possible mental state interpretations (Baker, Saxe, & Tenenbaum, 2009), and ToM reasoning might be qualitatively different from other kinds of reasoning.

Some studies have shown similar but uncorrelated developmental trends in ToM tasks and non-mental tasks that require similar representations (Arslan, Hohenberger, & Verbrugge, 2012; Flobbe, Verbrugge, Hendriks, & Krämer, 2008). For example, a relative clause in the sentence "The goat that pushes the cat" requires a similar representation as the complement clause in "Alice knows that Bob is writing", but only the complement clause requires a mental state representation. As children become older, they get better at understanding both types of sentences. However, their performance does not correlate when the factor age is controlled for. These findings show that ToM tasks might consume unique cognitive resources. It is important to note, however, that these tasks might have differed with respect to other factors, besides the aspect of mental representations.

Some studies show similar performance in ToM tasks, on the one hand, and equivalent but non-mental control tasks, on the other. In the false-belief or Sally-Anne task, for example, children have to attribute a false belief about an object's current location to Sally (Wellman et al., 2001; Wimmer & Perner, 1983). Sally stores an object at location A, but the object is moved from location A to location B while Sally is away. Therefore, Sally still thinks that the object is at location A. To pass this task, children should acknowledge that Sally falsely believes that the object is still at location A. The false-sign task is a similar but non-mental counterpart of the false-belief task. An object is first stored at location A, indicated by an arrow. Next, the object is moved from location A to location B, but the arrow still points at location A. The false sign in this task is the arrow pointing at location A, which is similar to Sally's false belief. Children's accuracy in both tasks is similar, and their performance correlates, even after correcting for age (Apperly, 2011; Leekam, Perner, Healey, & Sewell, 2008; Sabbagh, Xu, Carlson, Moses, & Kang, 2006). This finding implies that mental state reasoning might not qualitatively differ from other kinds of reasoning.

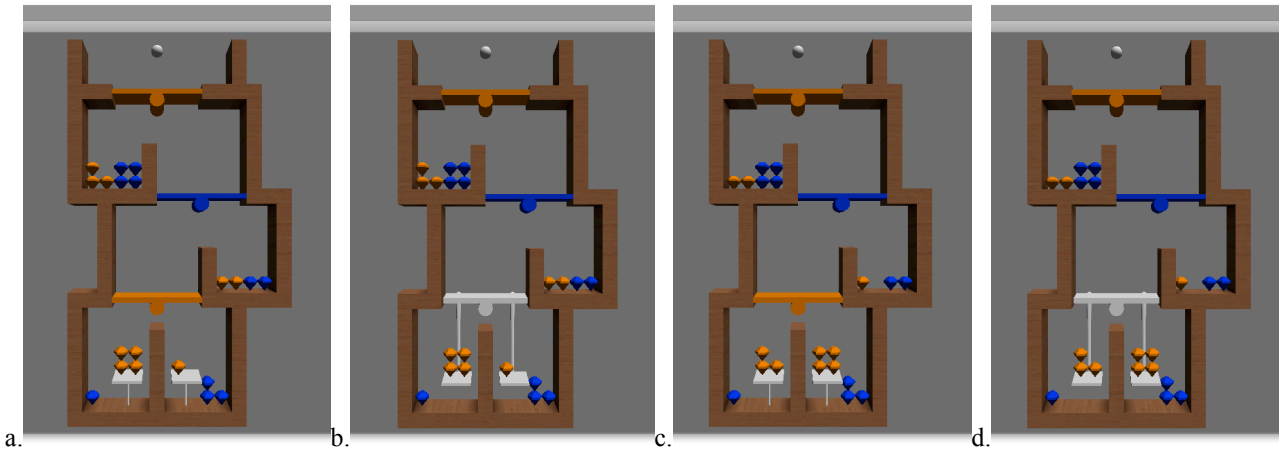

Figure 1. Examples of two-player Marble Drop games. A white marble is about to drop, and its path can be manipulated by turning the orange and blue trapdoors. In these example games, participants have to obtain as many orange diamonds as possible and they control the orange trapdoors. The other player has to obtain as many blue diamonds as possible and controls the blue trapdoor. In game *a*, the optimal decision for a participant is to let the white marble drop into the topmost bin, thereby obtaining 3 orange marbles. The 4 orange diamonds in the bottom-left bin are not obtainable, as the other (blue) player's optimal decision is to let the white marble drop into the middle bin: The other player knows that the optimal (orange) decision at the bottom trapdoors is to go left, yielding a suboptimal outcome of 1 blue diamond for Player 2. Games *a* and *c* are second-order games, because participants (as Player 1) have to reason about the other player (i.e., Player 2) who in turn has to reason about Player 1. The games in *b* and *d* are first-order counterparts of the games in *a* and c, respectively. They require the same comparisons, as the outcome of the balance scale is congruent with Player 1's last correct / rational decision: Both depend only on Player 1's diamonds in the bottom two bins. However, the games with the balance require one fewer switch between Player 1 and Player 2 perspectives.

Similar accuracy of responses in ToM tasks and their non-mental counterparts, however, does not necessarily imply a similar reasoning process. Moreover, differences might manifest themselves elsewhere, for example, in the reaction times. If, for example, both tasks require overlapping cognitive functions but ToM tasks require additional cognitive processing, the response patterns might not differ as much as the associated response times. Moreover, differences in accuracy might not manifest themselves until the tasks become more complex and exhaust cognitive resources.

Given these mixed findings, the question remains whether reasoning about mental states requires additional cognitive resources. Complex reasoning tasks consume cognitive resources, because oftentimes they require integration of information in the overall reasoning process. Integration of information and reasoning require executive functions such as planning, set shifting, resistance to interference, and working memory. It is not yet obvious why these executive functions alone would not suffice to reason about mental states.

In this study we investigate whether reasoning about mental states consumes unique cognitive resources. Participants are presented so-called Marble Drop games (Figure 1) in which they have to reason about another player. Marble Drop games have a recursive structure because the best possible, or optimal, decision at the first trapdoor depends on the other player's decision at the second trapdoor, which in turn depends on the outcome at the third trapdoor (Meijering, Van Rijn, Taatgen, & Verbrugge, 2011). The crucial factor in this experiment is whether the outcome at the third trapdoor is determined by Player 1's decision (*player* condition) or by the physics of a balance scale (*balance* condition). Both conditions require the same comparisons, but games in the player condition require one additional switch between player perspectives: Player 1 has to reason about what Player 2 thinks that Player 1 will do at the final trapdoor. If reasoning about mental states requires additional cognitive resources, games in the player condition would be more difficult than games in the balance condition.

Method

Participants are always assigned to the role of Player 1, and in both conditions they need to take the perspective of Player 2 to predict the outcome at the second trapdoor. This perspective taking requires ToM. As explained previously, the decision at the second trapdoor depends on the outcome at the third trapdoor. If the participants (i.e., Player 1) control that trapdoor, they need to switch perspective again. They need to re-take their own perspective from within Player 2's perspective. This requires second-order ToM. In the balance scale condition, participants do not have to switch perspective again, and thus need first-order ToM at most. They still need to make the same comparisons, as the outcome of the balance scale depends on Player 1's payoffs

and this outcome is congruent with Player 1's goal to maximize his or her payoffs.

If ToM requires unique cognitive resources, we expect that participants respond faster in the balance condition than in the player condition, because the balance condition requires one switch less between Player 1 and Player 2 perspectives than the player condition. We also expect better performance in the balance condition, because Marble Drop games in which Player 1 controls the third trapdoor might appear to be less deterministic. The assumption, here, is that it is easier to attribute knowledge of physics to Player 2 than to attribute to Player 2 epistemic reasoning about Player 1, as epistemic reasoning involves testing of multiple possible Player 2 perspectives.

Participants

Forty-two first-year Psychology students (30 female) participated in exchange for course credit. The average age was 21 years, ranging from 18 to 25. Each participant reported normal or corrected-to-normal visual acuity.

Stimuli

Of all possible payoff structures, only those that are diagnostic of second-order ToM reasoning were included in the experiment. A game is diagnostic of second-order ToM reasoning if it requires a participant to reason about each decision point to arrive at the optimal decision. An example of a non-diagnostic payoff structure is one in which Player 1's first payoff, in the topmost bin, is the maximum payoff in that game. In that case, Player 1 would not need to reason about the second and third decision points. The payoff structures are listed in a table, which can be found at http://www.ai.rug.nl/~meijering/marble_drop.html.

Design

The experimental design consists of two between-subjects conditions: balance condition versus player condition. In the player condition, participants are presented with the original second-order ToM games (Meijering, Van Rijn, Taatgen, & Verbrugge, 2012). In the balance condition, participants play the games with the same payoff structures, but the third decision point is replaced by a balance scale. Importantly, the games in both conditions are equivalent, as they require the same comparisons between payoffs. In each game, the outcome of the balance is the same as the last correct / rational decision of Player 1, because both only depend on the number of Player 1 diamonds in the bottom two bins (see Figure 1).

Procedure

After giving informed consent, participants were seated in front of a 24-inch iMac. They were randomly assigned to the balance scale condition or the player condition. The participants were instructed that their goal was to obtain as many diamonds as possible of their target color, either blue or orange, which was counterbalanced between participants. They were also instructed that Player 2's (i.e., the computer's) goal was to obtain as many marbles as possible of the other color.

The experimental procedure is the same in both ToM conditions. Participants are presented 62 unique games. At the start of each game, participants have to decide whether to stop the game, by letting the white marble drop into the top bin, or to continue the game, by letting the white marble drop onto Player 2's trapdoor. The game stops if Player 2 decides to let the white marble drop into the middle bin. If Player 2 decides to let the white marble drop onto the third trapdoor, participants in the player condition have to decide whether to stop the game in the bottom-left or bottom-right bin. In the balance condition, the physics of the balance scale determine whether the marble drops into the bottom-left bin or the bottom-right bin. Importantly, the balance scale is set in motion as soon as the white marble drops onto it. Otherwise, Player 2 would not have to reason about the balance scale. Each game is fully animated. See Figure 1 for some example games.

After each game, participants receive feedback that mentions Player 1's outcome. If, for example, the marble drops into a bin that contains two diamonds for Player 1, the feedback mentions: *"You get 2"*.

To familiarize participants with the rules of Marble Drop games, participants are presented additional feedback during the first 12 games. Feedback explicitly mentions whether the outcome is the highest attainable Player 1 payoff. In case a participant obtains 3 diamonds and could not have obtained more, feedback is: *"Correct. You get 3. The highest possible payoff!"*. In case a participant obtains 3 diamonds, but could have obtained 4, feedback is: *"Incorrect. You get 3. You could have obtained 4"*.

Results & Discussion

The data consist of 62 unique Marble Drop games (i.e., payoff structures) for each participant. In the statistical analyses, the games are blocked to accommodate non-linear and differential learning rates: The first 12 'training' games comprise the first block, and the remaining 50 games are split into 5 subsequent blocks of 10 games each. The graphs show means and standard errors, which are represented by error bars.

The data are analyzed by means of *linear mixed-effects models* (Baayen, 2008; Baayen, Davidson, & Bates, 2008) to accommodate random sources of variation due to sampling of participants and items (i.e., payoff structures). Specifically, each model allows for by-participant and by-item adjustments of the intercept. For each analysis that we report below, we first constructed a full factorial model with all main and interaction effects. Based on likelihood ratio comparisons, we removed main and interaction effects for as long as the corresponding parameters were not justified. If a comparison preferred a simplified model, we report the log-likelihood statistics. The correctness of responses is analyzed by means of *logistic* linear mixed-effects models, as correctness of responses is a binary variable (incorrect vs. correct).

Mean Proportion Correct

The proportion of correct responses in each block is averaged across participants and depicted in Figure 2. The figure does not show great differences between performance in the balance scale and player conditions.

A full-factorial model with main effects and an interaction effect of *Condition* and *Block* did not fit the data better than an additive model, $\chi^2(5) = 6.08$, *ns*. The parameters of the additive model are discussed below.

There is a significant effect of *Block*, $\beta = 1.37$, $z = 10.89$, $p < .001$. As can be seen in Figure 2, performance increases over the course of playing many Marble Drop games.

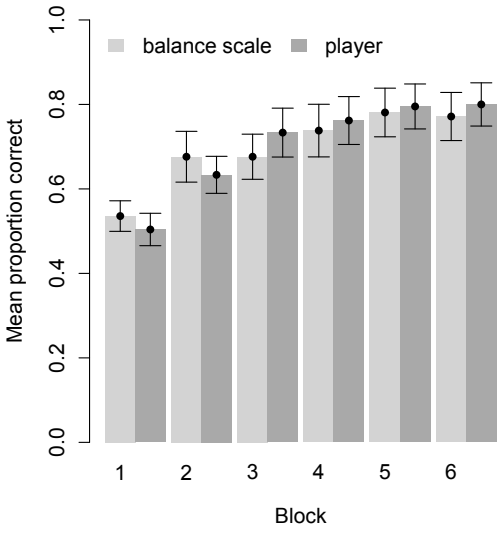

Figure 2: Mean proportion of correct responses per block; depicted separately for participants in the balance condition (light gray) and the player condition (dark gray).

There is no effect of *Condition*, as can be seen in Figure 2. In contrast to our hypothesis, the probability of making a correct decision does not differ between the balance scale and player conditions. An analysis of the types of errors (next section: Types of Errors), however, shows differential errors between the balance scale and player conditions.

Types of Errors

The errors that participants made were categorized according to game type, as an overall analysis might not be sensitive enough to differentiate between the balance scale and player conditions. Two types of games were distinguished on the basis of Player 2's (programmed) decision, which is either stop the game or continue.

There is no main effect of *Player 2 decision*, $\beta = -.08$, $z = -.575$, *ns*, which means that the difficulty of a game does not depend on Player 2's decision. This finding implies that there is no reason to believe that there are particular subsets of hard(er) payoff structures among the selected payoff structures.

There is a significant interaction effect between the factors *Condition* and *Player 2 response* (see Figure 3), $\beta = .65$, $z = 3.349$, and $p < 0.001$. In the balance scale condition, the probability of making a correct decision does not differ between games in which Player 2's decision is to stop, on the one hand, and games in which Player 2's decision is to continue, on the other hand. In the player condition, in contrast, there is a difference. One possible explanation is that participants in the player condition expect Player 2 to continue in most games, and this expectation pays off in games in which Player 2 actually decides to continue. In each game, Player 2 has a greater payoff in one of the last two end states than in the earlier end state, and participants might assign too great a probability to Player 2 going for that payoff. Participants in the balance condition, in contrast, might estimate those probabilities more accurately (i.e., *lower*), because games with a balance scale can be considered more deterministic.

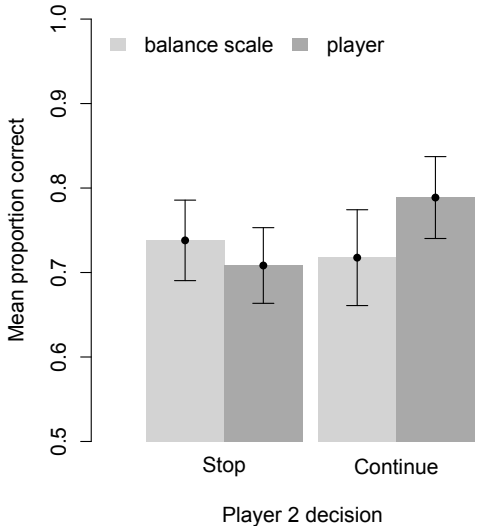

Figure 3: Mean proportion of correct responses across participants, depicted separately for the balance scale and player conditions, and Player 2's decision.

Reaction Times

There are differences in the types of errors between participants in the balance scale and player conditions, but what about the reaction time data? RTs are analyzed to find out whether a switch between perspectives comes with a time-cost. The RTs are log-transformed as reaction times are skewed to the right. Figure 4 shows the average log-RT across participants.

Figure 4 shows differential learning rates between participants in the balance scale and player conditions, especially in the first half of the experiment, in blocks 1 to 3. In the second half, blocks 4 to 6, the learning rates do not seem to differ that much. To specifically accommodate for differential learning rates, the factor *Block* was re-parameterized as a new factor *Half*, with levels *1* and *2*, and

a new factor *Block* with levels 1, 2, and 3 within each level of *Half*. The results of the full factorial LME with main and interaction effects of *Condition*, *Half*, and *Block* are discussed below.

The main effects of *Half* and *Block* (with linear contrast) are significant, $\beta = -.22$, $t = -7.82$, $p < .001$, and $\beta = -.18$, $t = -5.37$, $p < .001$, respectively. From the first to the second half of the experiment, and within each half, the RTs decrease linearly. The interaction between *Half* and *Block* is also significant, $\beta = .15$, $t = 3.19$, $p = .0015$. The decrease in RTs is stronger in the first half of the experiment than in the second half.

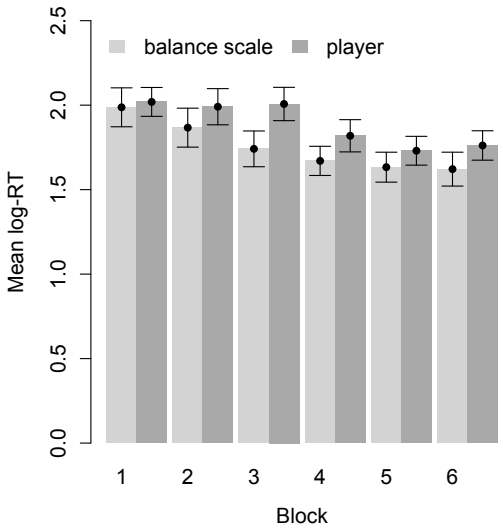

Figure 4: Average log-RT across participants plotted against block, separately for the balance scale and player conditions.

The interaction between *Condition* and *Block* is significant, $\beta = .17$, $t = -3.43$, $p < .001$. The decrease in RTs in the first half of the experiment is less strong in player condition than in balance scale condition. This finding is partly congruent with the hypothesis that RTs are shortest in the balance scale condition because it requires fewer switches between perspectives than the player condition. There is, however, no main effect of *Condition*, $\beta = .14$, $t = 1.2$, *ns*. Thus, on average, the RTs do not differ between the balance scale condition and the player condition. However, participants in the balance scale condition *do* become faster towards the end of the first half of the experiment, whereas participants in the player condition do not become faster. A possible explanation is that participants in the balance scale condition are quicker over the course of playing multiple games to attribute an understanding of gravity to Player 2. In contrast, participants in the player condition need to play more games and test multiple Player 2 perspectives.

The interaction between *Condition*, *Half*, and *Block* is also significant, $\beta = -.14$, $t = -2.83$, $p < .005$. As can be seen in Figure 4, the differential learning rates in the first half of the experiment disappear in the second half of the experiment, where the RT trends do not differ that much between the balance scale condition and the player condition.

In sum, there is an interaction effect of *Condition* and *Block* on the RTs, and this effect is mainly present in the first half of the experiment. There, the RTs decrease more in the balance scale condition than in the player condition. This interaction effect, between *Condition* and *Block*, seems to disappear in the second half of the experiment. A possible explanation for the latter finding is that, initially, participants in the balance scale condition settle more quickly on the correct Player 2 perspective than participants in the player condition, who test multiple Player 2 perspectives across multiple games.

General Conclusions

In this study we investigated whether ToM requires additional cognitive resources. We presented two types of games that required the same comparisons but differed with respect to the required depth of ToM reasoning: Games in the *player* condition required second-order ToM, as participants had to reason about a Player 2 that, in turn, reasoned about them; Games in the *balance scale* condition required first-order ToM, as participants had to reason about a Player 2 that reasoned about a balance scale. Our results show different errors between these conditions, which implies that the reasoning was not the same in the balance scale and player conditions. Moreover, the reaction time trends differed. The learning rate was faster for participants in the balance scale condition than for participants in the player condition. A faster learning rate in the balance condition is congruent with our hypothesis that it is easier to play against a Player 2 that reasons about gravity than playing against a Player 2 that reasons about mental states.

We assumed that games with a balance scale are easier to play because they appear to be more deterministic than games in which Player 1 has the last decision. This assumption is congruent with the RT data: Longer RTs in the player condition could be the cause of participants' testing of multiple possible Player 2 perspectives. Games in the balance scale condition, in contrast, require testing of fewer possible Player 2 perspectives, yielding shorter decision times.

Besides a faster learning rate in the balance condition, we expected a greater proportion of correct decisions. However, the probability of making a correct decision does not differ between the balance scale (i.e., first-order ToM) condition and the player (i.e., second-order ToM) condition. One possible explanation is that knowledge about gravity is not automatically attributed to Player 2. We expected that participants in the balance condition would automatically 'see' how Player 2's decision depends on the outcome of the balance, as young children have already mastered many balance scale configurations (Van Rijn, Van Someren, & Van der Maas, 2003). However, attributing an

understanding of gravity to Player 2 might be less of an automatic process than reasoning about gravity oneself.

Based on our findings, we conclude that participants do need ToM in Marble Drop games. Sequential games such as Marble Drop can be critiqued for not requiring ToM: If Player 2's strategy is known, the optimal (Player 1) decision can be determined without reasoning about Player 2's reasoning about Player 1's last possible decision. Applying backward induction, an algorithm based on sequential payoff comparisons, would yield the optimal decision. However, Meijering et al.'s (Meijering et al., 2012) eye tracking study shows that participants use more complicated and diverse reasoning strategies, not only backward induction. Moreover, backward induction would not be able to account for different types of mistakes and differential reaction times in the two conditions, as backward induction always works the same, irrespective of condition. Our findings provide strong support for the idea that sequential games are not just a decision-making problem but also evoke reasoning about mental states and thus require ToM.

In fact, it seems that sequential games are a particularly good paradigm to test reasoning about mental states, as they require *active* application of ToM. If Player 2's strategy is not yet known, participants need to actively find the correct Player 2 perspective. In any given game, multiple Player 2 perspectives might apply, but only that of a rational Player 2 is consistent with Player 2's actual decisions across all games. Active application of ToM is required to test multiple perspectives and find that of a rational Player 2.

To conclude, our findings are congruent with findings from fMRI studies showing that mental state reasoning employs brain regions that differ from the regions involved in cognitive control (Apperly, 2011; Saxe, Schulz, & Jiang, 2006). Our findings suggest that perspective taking requires additional cognitive resources, as opposed to just greater cognitive control, as one additional switch between perspectives induces not only longer reaction times but also qualitatively different decisions.

References

Apperly, I. A. (2011). *Mindreaders: The cognitive basis of "theory of mind."* Hove, UK: Psychology Press.

Arslan, B., Hohenberger, A., & Verbrugge, R. (2012). The development of second-order social cognition and its relation with complex language understanding and memory. In N. Miyake, D. Peebles, & R. P. Cooper (Eds.), *Proceedings of the 34th Annual Conference of the Cognitive Science Society* (pp. 1290–1295). Austin, TX: Cognitive Science Society.

Baayen, R. H. (2008). *Analyzing Linguistic Data: A Practical Introduction to Statistics Using R*. New York, USA: Cambridge University Press.

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

Baker, C., Saxe, R., & Tenenbaum, J. B. (2009). Action understanding as inverse planning. *Cognition*, *113*(3), 329–349.

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. (1992). Why the child's theory of mind really is a theory. *Mind and Language*, *7*(1-2), 145–171.

Leekam, S., Perner, J., Healey, L., & Sewell, C. (2008). False signs and the non-specificity of theory of mind: Evidence that preschoolers have general difficulties in understanding representations. *British Journal of Developmental Psychology*, *26*(4), 485–497.

Leslie, A. M., Friedman, O., & German, T. P. (2004). Core mechanisms in "theory of mind." *Trends in Cognitive Sciences*, *8*(12), 528–533.

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

Meijering, B., Van Rijn, H., Taatgen, N., & Verbrugge, R. (2011). I do know what you think I think: Second-order theory of mind in strategic games is not that difficult. In L. Carslon, C. Hoelscher, & T. F. Shipley (Eds.), *Proceedings of the 33rd Annual Conference of the Cognitive Science Society* (pp. 2486–2491). Austin, TX: Cognitive Science Society.

Onishi, K. H., & Baillargeon, R. (2005). Do 15-month-old infants understand false beliefs? *Science*, *308*(5719), 255–258.

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

Sabbagh, M. A., Xu, F., Carlson, S. M., Moses, L. J., & Kang, L. (2006). The development of executive functioning and theory of mind: A comparison of Chinese and U.S. preschoolers. *Psychological Science*, *17*(1), 74–81.

Saxe, R., Schulz, L. E., & Jiang, Y. V. (2006). Reading minds versus following rules: Dissociating theory of mind and executive control in the brain. *Social Neuroscience*, *1*(3-4), 284–298.

Van Rij, J., Van Rijn, H., & Hendriks, P. (2010). Cognitive architectures and language acquisition: A case study in pronoun comprehension. Journal of Child Language. *Journal of Child Language*, *37*(3), 731–766.

Van Rijn, H., Van Someren, M., & Van der Maas, H. L. J. (2003). Modeling developmental transitions on the balance scale task. *Cognitive Science*, *27*(2), 227–257.

Wellman, H. M., Cross, D., & Watson, J. (2001). Meta-analysis of theory-of-mind development: The truth about false belief. *Child Development*, *72*(3), 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*(1), 103–128.