Title

Modeling How Inhibitory Control Affects False Belief Reasoning Over Time Using Dense Longitudinal Data

Journal

Proceedings of the Annual Meeting of the Cognitive Science Society, 48(0)

Authors

Poyraz, Elif Nur
Leslie, Alan M.
Hemmer, Pernille

Publication Date

2026

Modeling How Inhibitory Control Affects False Belief Reasoning Over Time Using Dense Longitudinal Data

Elif Poyraz (elif.poyraz@rutgers.edu)[1,2], **Pernille Hemmer**[1,2] **& Alan M. Leslie**[1,2]
[1]Center for Cognitive Science, Rutgers University
[2]Department of Psychology, Rutgers University

Abstract

An important puzzle in theory of mind development is the performance shift in false belief understanding between the ages of 3 to 4. Theory of Mind Mechanism (ToMM) theory (Leslie et al., 2005) argues that changes in inhibitory power (IP) accounts for this shift. Wang and colleagues (2019) developed a computational model to implement this theory, where they quantified the role of IP using group averaged data. Here, we extend the Wang and colleagues' (2019) model to account for changes happening across time at the individual level. We model a novel dataset where 14 children were repeatedly tested during this transitional period on average every 20 days. Their performance in high and low inhibitory demand false belief tasks (FBT), and an independent measure of their inhibitory control was collected. We find that the model framework is successful in predicting matching inferences on children's IP given their performance in different FBTs.

Keywords: Theory of Mind; Bayesian; Computational Model; Development; Inhibition; Longitudinal; Microgenetic

Introduction

The ability to attribute mental states such as desires, goals, and beliefs to others is referred to as *theory of mind* (ToM) (Premack & Woodruff, 1978), and is shown to improve dramatically between the ages of 3 and 4 (Wellman et al., 2001). Specifically, reasoning about the false beliefs of others has been seen as the litmus test for attributing theory of mind to an entity. When reasoning about another mind, a reasoner makes inferences based on numerous aspects such as how their informational access to the environment might affect that person's knowledge and belief states. Furthermore, if one assumes that people are rational actors, the causal chain between mind and behavior can be reversed in order to infer a mental state given a certain observed behavior (Baker et al., 2009). The main question for the current work is the well-known "3-to-4 shift" in reasoning about others' false beliefs (Wellman et al., 2001).

The standard false-belief task (FBT) is a commonly used measure of ToM (Baron-Cohen et al., 1985). In this task, Sally has a marble and she hides it in a box. Afterwards, she leaves the room. While Sally is gone, Anne takes the marble out of the box and hides it in the basket. Children are then asked this question: "When Sally comes back, where will she look for her marble?" If children can indeed represent false beliefs and attribute a false belief to Sally, then they should also be able to predict Sally's consequent behavior based on that false belief. A persistent finding obtained using this paradigm is that, while the majority of the typically developing children older than 4 are successful in predicting "Sally will look for the marble in the box.", majority of the children younger than 4 respond with the current location of the object and report that "Sally will look for the marble in the basket." (Wellman et al., 2001).

There are divergent theories explaining the developmental data (Gopnik & Astington, 1988; Gopnik & Wellman, 1992; Perner, 1991). One view, which underlies the modeling framework used in the current paper, proposes that there are limitations in the processing resources that younger children have, and these factors can explain the performance differences between 3-and 4-year-olds (Leslie & Polizzi, 1998; Leslie et al., 2005). Leslie and colleagues' model of Theory of Mind Mechanism (ToMM) + Selection Process (SP) postulates that in order to be successful in the false belief tasks, children, as well as adults, not only need to be able to reason about beliefs of others, but they also need to successfully select between multiple belief candidates to make the correct false belief attribution. Specifically, Leslie and colleagues (2005) argue that children need sufficient *inhibitory control* power to overcome their own, potent representation of the true state of the world, which is that "the marble is in the basket". In the standard false belief scenario, ToMM spontaneously computes the True Belief (TB), based on the object's current location (W, i.e., World State), and the False Belief (FB) based on the visual/informational access (V) of the person to the scene. By default, TB gets selected based on the assumption that beliefs ought to be, and generally are true (Dennett, 1989). Crucially, in order to be successful in the standard task, successful *inhibition* of the TB is necessary, so that the correct belief candidate, FB, can be *selected* as a response. Consequently, in tasks where the saliency of the TB is reduced by increasing the uncertainty of one's own representation of the true state of the world, less inhibition will be required to suppress the TB representation. For ease of comparison, we will refer to such tasks with reduced saliency of the TB as *Low Inhibitory Demand FBT*, and to the standard FBT as *High Inhibitory Demand FBT* in the remainder of the paper.

ToMM + SP model is not only supported by evidence (Friedman & Leslie, 2004a, 2004b, 2005; Setoh et al., 2016) but also was the theoretical basis for the computational implementation in Wang et al. (2019), and Poyraz et al. (2020). Wang and colleagues (2019) used behavioral data from both Low

In G. Boleda, I. Dautriche, M. Maldonado, C. Saldana and T. Torrent (Eds.), *Proceedings of the 48th Annual Conference of the Cognitive Science Society*

(LD) and High Inhibitory Demand (HD) false belief tasks to infer three- and four-year-olds' inhibitory power, capturing inhibition's role in reasoning about false beliefs for the two age groups at an aggregate level. Poyraz and colleagues (2020) implemented the model using longitudinal single case data obtained in Baker and colleagues (2016), where individual children were tested repeatedly with the High Demand FBT. They were able to account for differences in individual children's performance in a longitudinal timescale (Poyraz et al., 2020). The current work expands on the two former implementations by modeling a new and unique dataset that captures performance in the LD and HD false belief tasks, as well as an independent measure of inhibition, in the longitudinal timescale. In the next section, we turn to the distinction between Low and High Demand FBTs, and provide the details for the longitudinal dataset used in the current work.

Methods

The Data Set

The dataset for this model consists of 14 individual cumulative records in which each child was repeatedly tested across multiple sessions at set intervals (Poyraz, 2025, doctoral dissertation). These longitudinal single-case studies aimed to examine preschool children during the transitional period of theory of mind development through repeated testing over extended periods of time. In each session, children completed various Theory of Mind and Inhibition measures. The current modeling implementation focuses on a subset of tasks: the Low- and High-Demand FBTs and an inhibition measure, the Opposites Task. Further task details are provided in the Tasks section.

The longitudinal single-case method conceptualizes each child as an individual experiment and examines variability and stability in performance at the individual level using Bayesian Change-Point Analysis (Baker et al., 2016). Although the current dataset was also analyzed using this method, the scope of the present paper is limited to reporting performance on the Low- and High-Demand FBTs, the Opposites Task, and the resulting model implementation outcomes. Importantly, the data were not analyzed at an aggregate or group-average level, as the focus was on individual performance trajectories and model estimations, which were qualitatively evaluated through visual examination of the curves.

The preschoolers' age at first testing session ranged from 38 months to 55 months, with an average of 45 months (*Std* = 4.8 months). Each testing session had no more than six weeks in between and occurred every 20 days on average, with the exception of one individual record, who had approximately 40 days between testing sessions. On average, children completed 22.5 testing sessions (*Std* = 8.1), with a range of 9 to 33 sessions.

Tasks

High Demand (Standard) False Belief Task. In this story, there are two agents, Jack and Mary. Jack hides a basketball in a bucket, then goes back to his home. While Jack is gone, Mary moves the basketball from the bucket to a bag. After both agents have left the scene, the children answered two control questions in the following fixed order: "At the beginning of the story, where did Jack put the ball?" and "Where is the basketball now?" The children were given feedback to the control questions. If they answered correctly, the experimenter said, "That's right!" If they answered incorrectly to one or both control questions, the story was repeated in full once more before asking the test question. The test question was asked at the very end, "Jack is going to come back now, and he wants to get the ball. Where will Jack look for the basketball?" No feedback was given to the test question. The children received a pass/fail (i.e., 1/0) score.

Low Demand False Belief Task. The low demand task followed a similar structure to the HD task, with the exception that instead of hiding an object, the agent hid an animal, which left the scene on its own at the end of the story. The children saw an agent, Maya, at the beach. Maya hides a crab in a bucket, then goes swimming. While Maya is gone, the crab comes out of the bucket and hides in the beach bag. Then, the crab comes out of the beach bag and leaves the scene. The crab is described as "going far far away, no one knows where the crab went!" This modification is known to reduce the saliency of the representation that the children themselves hold about the animal's last location, therefore reducing their certainty. In return, this is assumed to require less inhibition to suppress their own belief about where the object is. Children were asked the same kind of control questions as the High Demand task, with the same passing criteria. No feedback was given to the test question. The children received a pass/fail (i.e., 1/0) score.

Across sessions, children were presented with different versions of the FBTs where the character looks, names and the story set up was unique. The essential skeleton of the task structure was exactly the same across sessions.

Inhibition Task: The Opposites Task. This inhibition task was developed by Baker and colleagues (2010). On every trial, children are presented with a picture pair either showing semantically opposing pairs (e.g., full-empty jugs) or varying pictures along a similar dimension (e.g., a flying vs. a swimming duck). Each trial experimenter would say out loud one of the picture's name (e.g., "empty"). In the incongruent block, children are instructed to follow a "very silly rule" and point to the opposite picture of what the experimenter says. In the congruent block, children simply need to point to what the experimenter says. Children received 16 trials in total, 8 in each block. The first two trials of each block were practice trials, where children received feedback. After the practice trials, children are told that they "will play the game for real now". During the 6 experimental trials in each block, children did not receive any feedback. Children completed the incongruent block first followed by the congruent block. Children's first point was coded as their response. They received a score out of 6 for each block. For the modeling purposes of the current paper, children's performance in the Incongruent Block is used as a measure of their Inhibitory Control. Children received the same picture pairs across sessions, but the order of presentation of pairs and the word that the experimenter said for each word pair (e.g., saying "full" or "empty") was randomly selected across sessions for any given picture pair.

The Formal Model

Model Overview

Following convention, we adopt graphic models as a way to visualize structure. In the graphic model, circular nodes represent all continuous variables. Nodes for observed variables are shaded while nodes for latent variables are unshaded. Further, stochastic variables are differentiated with a single border and deterministic variables with double borders. Arrows indicate conditional dependencies. Figure 1 shows the graphical model implemented in the current work, which assumes that I for each child is allowed to vary over a temporal window. The figure displays the model details for both the HD and LD tasks, where the additional uncertainty parameter highlighted with the red dashed square is added when modeling the LD task.

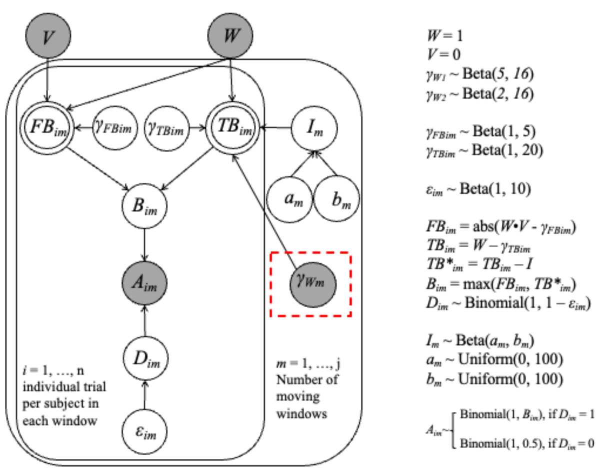


Figure 1: The graphical model for ToMM+SP for a High IC Demand task. The red dotted parameter is added for modeling the Low IC Demand task. W: world, FB: false belief; TB: true belief; D: desire; A: action; V: visual access; I: inhibition. The model assumes that I for each child is allowed to vary over a moving temporal window.

Using the longitudinal data set, we implement a 'moving window' that iterates over a set number of trials per subject. A depiction of how this window was implemented on data can be found in Figure 2. This window size was set to five for the current model, indicating that for each inference, the model received five data points for A. The criterion for the least amount of trials that one could expect a change to be manifested is largely unknown. Following Poyraz et al. (2020), having the window size set to five seemed like a reasonable number that could capture change given that each session is separated by no more than 6 weeks.

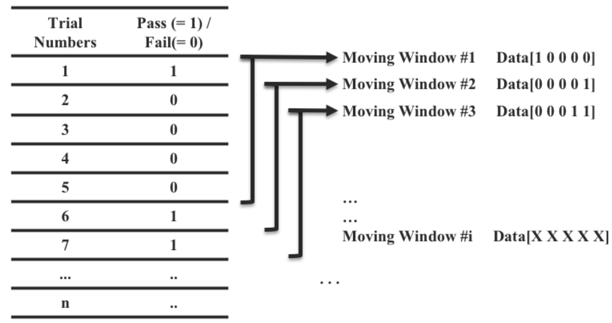


Figure 2: Illustration of how data was selected by the use of a moving window of 5 trials.

Importantly, children in this dataset completed a High Demand (standard) and a Low Demand FBT, as well as an independent measure of inhibition. Therefore, we can model this longitudinal dataset with both the High Demand and the Low Demand version of the model, implementing the Temporal aspect of the data again with the use of moving windows. Furthermore, we can compare the inferred Inhibition values of the model across the two tasks against the actual observed inhibition performance.

Parameter Specifications

In the false-belief task, Sally wants her toy back,and will go to the location where she believes the toy to be. Thus, Desire (D) is assumed to have a large prior probability, $1-\varepsilon$. Following Wang et al. (2019), we;

1. use an asymmetric beta prior on ε, with $\varepsilon \sim \text{Beta}(1, 10)$, indicating that Sally is unlikely to change her desire (this is also the assumption used by Goodman et al., 2006);
2. assume the TB content one attributes to Sally is determined by one's own knowledge of the world with probability $W - \gamma_{TB}$ ($W = 1$ is change in the world or 0 is no change) that reflects the variations in how one may incorrectly represent the world. The uncertainty introduced by the low demand task is also incorporated with an additional parameter γ_W, which takes an asymmetric beta prior of $\gamma_W \sim \text{Beta}(5,16)$ for children who were 3-year-old at the beginning of testing, and $\gamma_W \sim \text{Beta}(2,16)$ for 4-year-olds (Wang et al., 2019);
3. assume the FB content is determined by Sally's visual access to the world (V=0, or no access, in a false-belief scenario), with the uncertainty about FB given by the absolute value of $W \cdot V - \gamma_{FB}$. Following Goodman et al., we used asymmetric beta priors on γ_{TB} and γ_{FB}, with $\gamma_{TB} \sim \text{Beta}(1, 20)$, and $\gamma_{FB} \sim \text{Beta}(1, 5)$, indicating that one has an accurate representation of the world, but the calculation of others' visual access may be prone to more variation therefore affecting the precision of the FB representation.

Under the assumption of ToMM + SP, to succeed in the standard false-belief task, one has to deploy sufficient inhibition to overcome the TB prior. Following Wang et al., $I \sim \text{Beta}(\alpha, \beta)$, where $\alpha \sim$ Uniform(1,100) and $\beta \sim$ Uniform(1,100), indicating no prior assumptions about the shape of the beta distribution of I and to allow for a complete estimation of I from the observed data.

The attributed belief B is the stronger belief between FB and TB with inhibition, i.e., the probability of B is the larger of the probability of the true belief with an inhibition $W - \gamma_W - \gamma_{TB} - I$ and the probability of the FB (absolute value of $W \cdot V - \gamma_{FB}$. After the selection, $B \sim$ Binomial(1,max($FB, TB - I$)), one attributes either the belief that there is no change $B = 0$ or the belief that there is a change $B = 1$ to the agent, which coupled with desire determines action prediction A. If the attributed desire is to retrieve the object $D = 1$, prediction follows the attributed belief B and $A \sim$ Binomial(1, B); if the attributed desire is ambiguous $D = 0$, action prediction is random and $A \sim$ Binomial(1, 0.5). See Wang et al. (2019) for the conditional probabilities among all variables.

In sum, A is conditioned on the B, and D. B is determined by selecting between the TB and the FB through applying I.

The models were implemented using WinBUGS (Lunn et al., 2000; Ntzoufras, 2009). Our results are based on drawing 3,000 samples from three separate chains with a 100 burn-in period for each of the models. Convergence of the chains was assessed using the $\hat{\text{R}}$ statistic (Brooks and Gelman, 1998).

Results

Figure 3 shows the model's results for High and Low Demand frameworks. Each row represents a subject's model fit in three subplots. At the end of each row, the shaded boxes display the subject ID and their age in months at the start of testing.

1. The left column, *Observed Performance*, shows raw performance across moving windows for the High and Low Demand FBT.
2. The center column, *Action Prediction*, shows the mode of simulated Action prediction values based on inferred I estimates per window.
3. The right column, *Inferred Inhibition* shows the mode of the Posterior distribution of Inhibition per moving window. Additionally, the right column subplots include children's experimentally assessed performance on the independent Inhibition task, the Opposites Task, in a blue dashed line.
4. The three panels, marked by red, blue, and gray lines, represent groupings reflecting the degree of match or mismatch between performance on the Opposites Task and the model's inferences on inhibition. These groupings were determined through qualitative visual inspection of the curves.

The Y axes are at the same scale in all three subplots. The left and center columns show proportion accuracy and model-simulated proportion accuracy, so the Y-axis range is 0 to 1. The right plot shows the model's inhibition estimates which could range from 0 to 1, so the Y-axis range is also 0-1.

The X axes show the number of windows, which represents time passing or the developmental timescale. The number of windows differs across subjects as a function of the number of trials (number of windows = number of trials - window size + 1; where window size is set to 5).

Summary of Experimental Data

Children's observed performance on the two FBTs are quite variable (Fig.3, Left Column), with no clear pattern based on chronological age. For example, Subjects 7 and 4 enrolled in the study at the same chronological age (i.e., at 38 months) and participated for comparable number of sessions, but their performance trajectories look drastically different. Interestingly, Subject 4's observed inhibition performance (Right Column, Blue Dashed line) fluctuates more, both when compared against Subject 7's inhibition, and most of the other subjects. Five children show the most amount of fluctuation in inhibition compared to the rest of the sample, Subjects 9, 5, 4, 3, and 1, whereas the performance in the two FBTs were more variable. Moreover, all children almost reached ceiling performance in inhibition by the end of testing. In fact, 10 out of 14 children come to ceiling performance very early on during testing. However, the same does not apply to their performance in the FBTs (e.g., Red Panel). We will return to the kind of implications this divergence has for the model implementation in the following section.

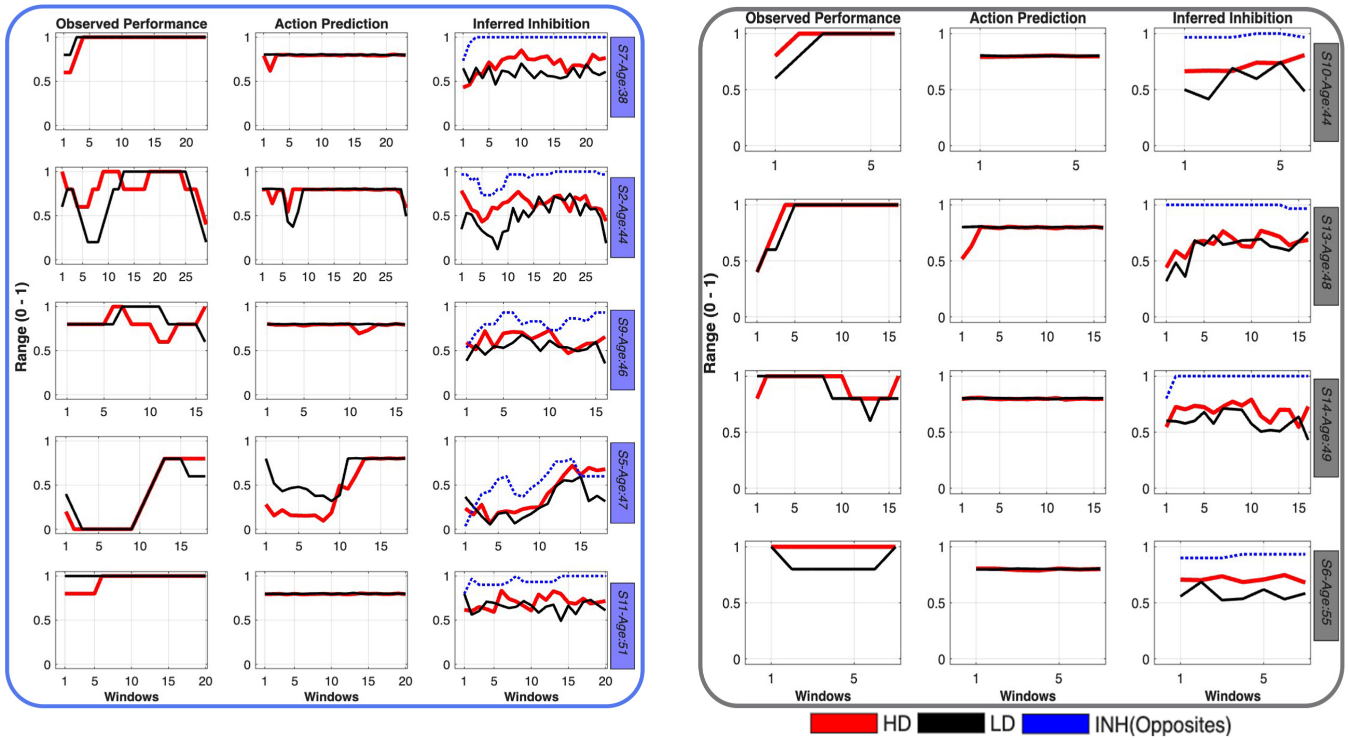
Model Findings

The model's findings offer unique insights because they use the same children's data within the same time frame for both tasks. This allows for three main comparisons. Firstly, one can compare the model inferences for High and Low Demand tasks across the Left and Center columns to check if the model simulations match the mean observed performances. In some cases, there are over- and under-estimations in Action prediction when observed performance is at the boundaries (i.e., either at the ceiling or the floor).

Secondly, by comparing the black and red curves in the Right column, one can observe if the inferred estimates for Inhibition from the model for the High and Low Demand FBTs match. Since the same children completed both tasks, the cognitive parameter I should not differ significantly between them. Indeed, the figures show that the two inhibition curves inferred from the High and Low Demand tasks closely follow one another across individuals, indicating the coherence of the ToMM+SP framework used to capture Inhibition's effect on the two versions of the false belief task.

Finally, we can compare the observed inhibition values (blue line) to the model inferences on I (red and black lines) in the Right subplots. Some children show deviations between model inferences and observed inhibition performance, while others have model inferences that closely follow their observed inhibition performance.

The Blue panel displays the children whose observed inhibition performance closely matched by the model inferences on I, namely Subjects 7, 2, 9, 5, and 11, sorted by an increasing order of age at the start of testing (Fig.3). The



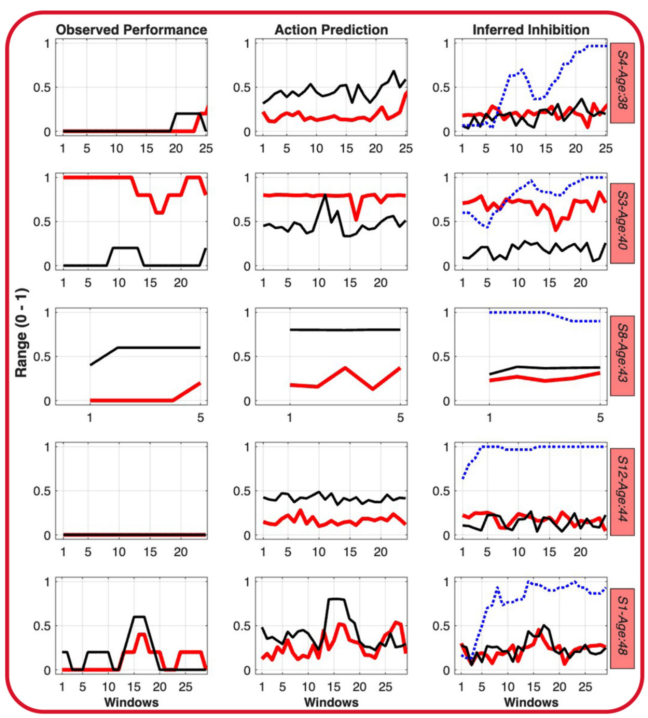


Figure 3: Each row shows a single subject's model fits across three columns. The shaded box at the end of each row reports the subject ID and age (in months) at the start of testing (e.g., S7-Age:38). Left column: Mean observed performance in the High Demand (HD) and Low Demand (LD) FB tasks (red and black, respectively). Center column: Mode of the model-simulated action prediction. Right column: Mode of the inferred posterior distribution of inhibition. The blue dashed line indicates mean observed inhibition performance in the Opposites Task. The x-axis represents the number of moving windows, and all y-axes range from 0 to 1. For the left and center columns, values indicate proportion correct; for the right column, values represent the inferred inhibition distribution over [0,1]. Subjects are categorized into three groups based on qualitative comparison of the right-column trajectories: blue panel, model estimates closely match observed inhibition; red panel, model estimates diverge from observed inhibition; gray panel, intermediate cases.

youngest, *Subject 7*, and the oldest, *Subject 11*, show almost at ceiling performance for both FBTs, as well as the inhibition task. Consequently, the model inferences - though a lot more variable- also portray matching inferences both in terms of simulated accuracy (Center column), and inferred inhibition (Right column). For the other 3 subjects in this group, we can observe that the performance of Subjects 2 and 9 fluctuated in the FBTs, and these fluctuations were well-captured by the model based on both model predictions (middle subplot) and inhibition inferences (Right column).

The Red Panel shows the group of children where the model inferences on I deviated from the observed inhibition performance; Subjects 4, 3, 8, 12, and 1 (Fig.3). Again in this group, we see that looking at chronological age alone is not sufficient for explaining the variances across individuals. For example, Subject 4 and Subject 12 both perform at floor for the FBTs though their age at the start of testing is 6 months apart, and they have completed similar number of sessions for a long duration. Moreover, this group of children for whom the model showed deviations in inhibition inferences vs. observed inhibition are the ones in the entire sample who performed poorly on the FBTs.

Across the board, the model inhibition inferences are estimated to be lower than children's experimentally assessed inhibition. This could be due to two main reasons, one about the experimental data, and the other about the model assumptions. The first reason, which is in the experimental data, we see that most of the children reach ceiling performance in the independent inhibition measure. This indicates either that children are getting better at this task as a function of their inhibitory control capacity increasing with age, or that they are getting better at the task itself which makes the task fall short of capturing the still-existing variance in their actual inhibitory control ability. The second reason is that the model's inferences on Inhibition are solely based on children's performance in the FBTs. As a result, model inferences capture variances in inhibition plus representing false beliefs. We will turn to discussing this point more in the Discussion section.

Discussion

Inhibition's role in ToM reasoning was previously captured with group-averaged data (Wang et al., 2019), accounting for differences in performance between three- and four-year-olds. We used the same model framework to model individual-level performances in FBTs over time, without added complexity. Using multiple measures of false belief, we tested the explanatory power of the ToMM + SP framework for children's individual ToM performance. We observed variable individual trajectories with occasional increases and decreases. Instead of relying on a crude age assumption, we incorporated Inhibition's effect to explain performance variances and changes. The model fit individual developmental curves well, capturing fluctuations in performance. Remarkably, the two inhibition curves inferred by the High and Low Demand tasks closely followed each other across individuals. This suggests that the ToMM + SP framework provides a cohesive and consistent framework for systematically capturing Inhibition's role in both Low and High Demand tasks. However, the model implementation at the individual level has room for improvement.

The model assumes that children's inhibitory control is captured solely by inferences from false-belief performance, which, in reality, is an oversimplification. The actual picture likely shows a significant overlap in performance between the two domains. These distinct mechanisms work independently, but their behavioral outcomes (i.e., performance) reflect a combination of both when reasoning about false beliefs. Thus, the false belief task may not fully capture inhibitory control. To capture nuances, one could compare actual inhibition *distributions* to those inferred in the model for each child over time. Achieving this would require more data points per child across all tasks, potentially over a shorter time span, to model inhibition with greater precision. Alternatively, increasing the size of the moving temporal window can increase variability, resulting in more data points per inference in *Inhibition*.

Another approach is to adjust the hyperparameters over the FB and TB nodes to restrict their distributions and reduce variance. The hyperparameters γ_{TB} and γ_{FB} were defined to capture the variance among individuals in the group-average data. However, the current implementation aims to model an individual's performance, which is expected to vary less than the variance between individuals. Adjusting these hyperparameters can improve the model fit for observed vs. inferred inhibition trajectories. Furthermore, the current implementation assumes the same inhibition distribution for different age groups, but it could be better to assume different distributions.

Importantly, the sequence where a child performs well in false belief but poorly in inhibition was not true for any cumulative record in the dataset. This shows that inhibition is *necessary* for either high performance or performance changes in the false belief tasks. One question remains: how much inhibition alone is *sufficient* in explaining performance variances? The current modeling work highlights the need to elucidate this gap.

Summary & Conclusion

The current work provided a model implementation using a unique dataset of cumulative records for individual children in a longitudinal change-point study (Poyraz, 2025). We examined how Inhibitory power varied across individuals in this dataset.

A key difference between the current work and Wang et al. (2019) is the nature of the data. We modeled each child separately, allowing us to infer different Inhibitory power values per case across two FBTs without adding parameters to the original model. Interestingly, in less than half of the cases, we found deviations between the model Inhibition estimates and the observed performance. Since Inhibition is an inferred estimate based on children's performances in the false belief tasks, these deviations are not surprising. Another significant difference is that the current work included multiple FBTs tested within subjects and an independent measure of inhibition. The diversity of tasks and the longitudinal timescale enabled novel interpretations.

Overall, this implementation highlights the immense variance in the actual developmental timescale across individuals, which cannot be explained by chronological age. It emphasizes the importance of collecting individual-level evidence to develop better theoretical and computational model frameworks.

Acknowledgments

The data and materials can be obtained upon request from the first author via email.

References

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

Baker, S., Friedman, O., & Leslie, A. (2010). The opposites task: Using general rules to test cognitive flexibility in preschoolers. *Journal of Cognition and Development*, *11*(2), 240–254.

Baker, S. T., Leslie, A. M., Gallistel, C. R., & Hood, B. M. (2016). Bayesian change-point analysis reveals developmental change in a classic theory of mind task. *Cognitive Psychology*, *91*, 124–149.

Baron-Cohen, S., Leslie, A. M., & Frith, U. (1985). Does the autistic child have a "theory of mind"? *Cognition*, *21*, 37–46.

Brooks, S., & Gelman, A. (1998). General methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics*, *7*, 434–455.

Dennett, D. (1989). *The intentional stance*. the MIT Press.

Friedman, O., & Leslie, A. M. (2004a). A developmental shift in processes underlying successful belief-desire reasoning. *Cognitive Science*, *28*, 963–977.

Friedman, O., & Leslie, A. M. (2004b). Mechanisms of belief-desire reasoning: Inhibition and bias. *Psychological Science*, *15*(8), 547–552.

Friedman, O., & Leslie, A. M. (2005). Processing demand in belief-desire reasoning: Inhibition or general difficulty? *Developmental Science*, *8*, 218–225.

Goodman, N. D., Baker, C. L., Bonawitz, E. B., Mansinghka, V. K., Gopnik, A., Wellman, H., & Tenenbaum, J. B. (2006). Intuitive theories of the mind: A rational approach to false belief. *Proceedings of the twenty-eighth Annual Conference of the Cognitive Science Society*.

Gopnik, A., & Astington, J. W. (1988). Children's understanding of representational change and its relation to the understanding of false belief and the appearance-reality distinction. *Child Development*, *59*, 26–37.

Gopnik, A., & Wellman, H. M. (1992). Why the child's theory of mind really is a theory. *Mind & Language*, *7*, 145–171.

Leslie, A. M. (1994a). Pretending and believing: Issues in the theory of tomm [(Reprinted in J. Mehler & S. Franck (Eds.), Cognition on Cognition, 193–220, MIT Press, 1995)]. *Cognition*, *50*, 211–238.

Leslie, A. M. (1994b). Tomm, toby, and agency: Core architecture and domain specificity. In L. Hirschfeld & S. Gelman (Eds.), *Mapping the mind: Domain specificity in cognition and culture* (pp. 119–148). Cambridge University Press.

Leslie, A. M., German, T. P., & Polizzi, P. (2005). Belief-desire reasoning as a process of selection. *Cognitive Psychology*, *50*(1), 45–85.

Leslie, A. M., & Polizzi, P. (1998). Inhibitory processing in the false belief task: Two conjectures. *Developmental Science*, *1*(2), 247–253.

Lunn, D. J., Thomas, A., Best, N., & Piegelhalter, D. (2000). Winbugs-a bayesian modelling framework: Concepts, structure, and extensibility. *Statistics and Computing*, *10*, 325–337.

Ntzoufras, I. (2009). *Bayesian modeling using winbugs*. Wiley.

Perner, J. (1991). *Understanding the representational mind*. MIT Press.

Poyraz, E. N., Hemmer, P., & Leslie, A. M. (2020). Can changes in inhibitory control explain child-level theory of mind development? *Proceedings of the forty-second Annual Conference of the Cognitive Science Society*, 1825–1832.

Poyraz, E. N. (2025). *The development of theory of mind and inhibition on the fly: Cross-sectional and longitudinal single-case studies, and a computational model* [Doctoral dissertation, Rutgers University]. ProQuest Dissertations & Theses.

Premack, D., & Woodruff, G. (1978). Does the chimpanzee have a theory of mind? *Behavioral and Brain Sciences*, *1*, 512–526.

Setoh, P., Scott, R., & Baillargeon, R. (2016). Two-and-a-half-year-olds succeed at a traditional false-belief task with reduced processing demands. *PNAS*, *113*(47), 13360–13365.

Wang, L., Hemmer, P., & Leslie, A. M. (2019). A bayesian framework for the development of belief-desire reasoning: Estimating inhibitory power. *Psychonomic Bulletin & Review*, *26*, 205–221.

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.