Longitudinal Multidimensional Item Response Modelling in Preschool Children's Mental State Understanding

by

Vilma Susana Romero Romero

Supervisor:

Dr. Gareth Ridall

Dissertation submitted in partial fulfillment for the degree of *Master of Science in Statistics*

September 2015

Abstract

Theory of Mind is the ability that allow an individual to understand what others believe, desire or think and to anticipate how they might react in a given situation based on that knowledge. This ability is known to be developed at around 4 to 5 years old and its acquisition is very important as it helps us to interact properly in the social environment. Researchers have been studying this topic since many years ago and there exist different mental states tasks to assess the ability in young children. In this study, 86 very young children (starting with ages around 32 months) were assessed with different mental state tasks three times over 4 months intervals in order to evaluate if there is an important manifestation and improvement of the ability for this range of age, which is over a year before they are supposed to pass the tests in this context. For this purpose, Multidimensional Item Response Theory was employed first to reduce the dimensionality concerning Theory of Mind. Then, each latent dimension found was evaluated under the Bayesian Longitudinal approach and the continuous unobservable abilities of the child at each time point were obtained. This last output helped us to build a causality diagram in which it can be shown how each ability is affected by all the others abilities at previous times. One of the main findings was that the construct of Theory of Mind can be comprised of 6 latent dimensions which are Non Verbal False Belief, Pretense, Desire and Think, Verbal False Belief, Deceptive Box, Narrative and Location Change. Moreover, it was found that Pretense, Desire and Non Verbal False Belief tasks were the abilities that evolved more in the study period of time. Regarding the causal analysis, it can be highlighted that Time 3 measures are best predicted by Desire, Pretense and Think measures at Time 2. Finally, the conclusions in the psychological context are discussed as well as the limitations and further work related to the methodology employed.

Keywords: Theory of Mind, Item Response Theory, Multidimensional Item Response Theory, Exploratory and Confirmatory Factor Analysis, Bayesian Longitudinal Analysis, Causal Analysis.

Acknowledgements:

First, I would like to thank my supervisor Gareth Ridall for all the support and advice throughout this dissertation. I really appreciate all the help provided to get the Bugs code run properly and all the final comments given to finish this amazing project. Also, I would like to thank Professor Charlie Lewis for all the explanation of Theory of Mind and the helpful comments given to understand the results according to the context, and to all the MSc. professors for all the valuable lectures given. Finally, I would like to thank my family, specially my parents, for their unconditional support and trust through this hard but rewarding year.

Contents

Chapter 1

Introduction

1.1 Overview

"Theory of mind" is the ability that allows individuals to perceive their own mental states as well as others', such as beliefs, desires and intentions. It also involves knowing that they differ from one person to another. According to psychologists, this ability is developed during the first years of life being the age of 4 (48 months), the crucial change in the development of theory of mind. Acquiring this ability is very important because children can understand the social environment and how to interact in it shaping their behaviour appropriately; as a result, these will make them build good relationships later in their lives (Wellman, 1990; Astington and Jenkins, 1995; Astington, 2001; Shakoor et al., 2012).

Different mental state tasks have been proposed to assess the acquisition of theory of mind in young children. Among those, the most common task is the so called false belief test, which has two formats: deceptive box and unexpected location change. The false belief tasks can also be verbal and non-verbal (Call and Tomasello, 1999). Another kind of tasks were introduced by Lillard to measure pretence, desire and think (Lillard and Flavell, 1992).

Many studies have been made in relation to the development and acquisition of the ability of theory of mind in preschoolers (children from 3 to 5 years old). Researchers have found an association between the performance in mental state tasks and family environments like birth of order, number of siblings, observation of interaction mother-sibling struggles (Dunn et al., 1991; Perner et al., 1994) and overall family size including other adult relatives at home (Lewis et al., 1996). Moreover, other studies have shown that theory of mind performance correlates with socio-cultural factors (Wellman et al., 2001; Rodrigues et al., 2015).

However, most of the literature involves qualitative studies and sometimes the performance of the mental state tasks is analysed with simple statistical techniques such as Pearson correlations, ANOVA, Mann-Withney tests or sometimes some more complex procedures like Logistic Regression. Therefore, there is a need to have a more complete analysis of the response patterns obtained from mental state tasks by applying appropriate sophisticated statistical techniques.

As explained before, theory of mind is a construct that it cannot be measured straightforwardly. By contrast, its evaluation relies on the responses of different mental state tasks delivered to children. Given the nature of the data obtained, which is binary due to the failure or success registered of the task, and the latent ability, the use of Item Response Theory procedure is suitable. This technique allows us to evaluate the interaction between children's ability with the difficulty of the item by mapping a probabilistic model for the pattern responses. Nonetheless, item response modelling relies on the assumption that the latent trait under study is unidimensional and sometimes this is not adequate because it can be a combination of nested factors within a main construct. Consequently, multidimensional item response modelling has to be taken into account to deal with this.

In the present study, 86 very young children (starting with ages around 32 months) were assessed with different mental state tasks three times over 4 months intervals. The aim of the study were to attempt demonstrate an understanding of mental states in children over the third year of life - that is over a year before they are supposed to pass belief tasks. In order to do this, first we identify underlying factors concerning Theory of Mind by applying Multidimensional Item Response Theory and then analyse each dimension under the Bayesian Longitudinal approach. An additional objective is to build a causal mechanism determination in order to discover how each ability is affected by all the others abilities at previous times. Finally, find out if the outcomes tight out with preceding Psychological theory.

The thesis is structured as follows: This chapter ends with a brief review of the literature. In Chapter 2 a short description of the psychological context is described as well as the mental tasks examined to the preschoolers. Chapter 3 shows an exploratory analysis of the data, including the issues concerning missing data, some non-longitudinal plots and a mixed logistic regression model applied for each item taking into account the existing variability among individuals. Chapter 4 will first present a general description of Item Response Theory and some simple models considered for binary data. This will be then extended to the Multidimensional design and will end with an application in the analysis of pattern responses to mental states tasks. Having identified the underlying dimensions of theory of mind tasks, a Longitudinal Item Analysis will be conducted in a Bayesian framework as well as a Causal Analysis in Chapter 5. Finally, the conclusions and future directions are stated in Chapter 6.

1.2 Literature Review

In this literature review, I will briefly explained some relevant previous studies related to the acquisition of theory of mind in young children.

One significant study comes from Wellman et al. (2001) who conducted a meta-analysis in which they compiled a broad range of false belief experiments which led to have under study about 5000 children. Two main results from this study was the fact that there is a clear transition in the acquisition of Theory of Mind at ages of 4 years and 4 months. Also, that there is variation between cultures. In order to get those outcomes, the authors employed initially a Logistic Regression technique comparing the relation between the proportion of correct responses and age as only factor. Later on, they introduced other independent variables in the regression like the type of task, question and culture as well as considered interactions between them.

Jenkins and Astington (2000) studied the relationship between children's theory of mind and their social behaviour by assessing them with different standard mental tasks and video tapping their behaviour at three different time periods. In order to have one measure of theory of mind, they summed all the scores obtained in each task (0 for failure and 1 for success). Moreover, they applied three repeated-measures analyses of variance (ANOVAs), one for each time point, to determine if there was a significant increase in children's theory of mind understanding, which led to a positive result. Taking only into account the part of theory of mind analysis, the methodology applied in this paper to obtain a total score seems fair, but in the process it looses information and it would be better if the pattern response is analysed instead. It would have also been better if the analysis of improving theory of mind was done within the longitudinal framework and not separately each time. However, the main goal of the paper was not the analysis of mental state responses but rather the correlation with social behaviour and for this the authors proposed two different global causal models through multivariate multiple regression, so the methodology of the total score looks sensible for this context.

A recent study conducted in Brazil highlights the relevance of socio-cultural factors in the development of theory of mind in preschool children (4 and 5 years old). Rodrigues et al. (2015) applied the Theory-of-Mind-Scale proposed by Wellman and Liu (2004) to assess the evolution of theory of mind considering interactions among gender, age and the kind of school they were from (public or private). To achieve this, Rodrigues and colleagues employed a multiple linear regression finding significant differences in age and the belonging to the type of school, but there were not differences regarding gender.

To sum up, it can be noticed that the literature regarding the analysis of only response patterns of mental state task is scarce or even null. Thus, it is important to analyse this with proper statistical techniques under the psychometrics context, in which Item Response Theory takes place.

Chapter 2

The Context

In this chapter, the psychological context of Theory of Mind is illustrated by giving a simple definition, followed by the explanation of commonly mental state tasks administered to children, which were used for the purpose of this thesis.

2.1 Theory of Mind

The best way to explain Theory of Mind is by stating an example. Imagine you are playing with your little sister of 3 years old. You told her a story about Laura playing with her doll, but suddenly she gets tired and goes to sleep leaving her doll in her basket. Her naughty cousin Amanda takes then the doll and put it into a box while Laura is sleeping. After a couple of hours, Laura wakes up and want to play with her doll, so you will ask your sister about where Laura will look for her doll. Even though, your sister is smart, she will probably answer that Laura will look in the box. This is because children around this age are not able yet to recognize that others do not have the same information as they do and that they have their own belief, even if this is wrong.

With the above example as a preamble, Theory of Mind regards the ability of a subject to understand what others believe, desire or think and to anticipate how they might react in a given situation based on that knowledge. Psychologists agree that as the child grows up, they are able to learn this ability being the age of 4 where suddenly becomes to be remarkable.

There are several tasks to evaluate the acquisition of this ability in very young children, but researchers mostly used the so called False Belief Task. In the following section, the most common mental state tasks in the context of False Belief are explained.

2.2 Mental State Tasks

(1) **Location Change**
This task is one of the pioneers among all the known tasks and it was proposed by Wimmer and Perner (1983). It was initially named "the Maxi Task" and basically consisted in a story structured as: Maxi puts his chocolate into a cupboard A and his mother takes it from cupboard A into cupboard B while he was absent. Then the children under study need to answer when Maxi will look for his chocolate when he

returns. A child who points out the location A was able to understand that others have beliefs and knowledge different of his own. Some years later, Baron-Cohen et al. (1985) modified this task and called it the "Sally-Anne Task", which procedure is explained in Fig. 2.1. Nowadays, this task is known as Location Change. Moreover, according to various studies, this task can be passed by children of 4 years 6 months.

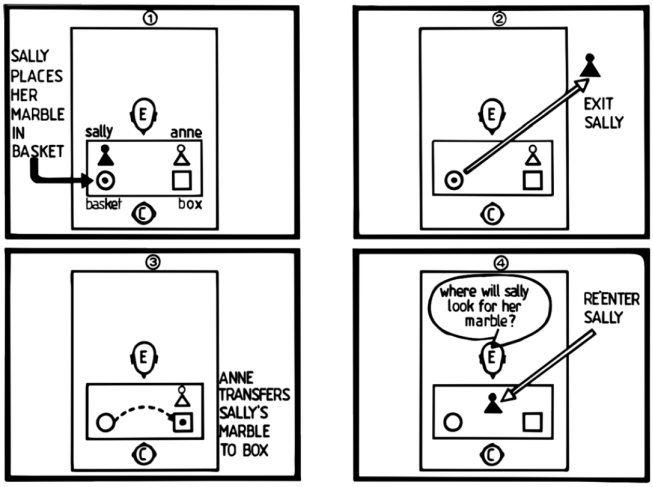

Figure 2.1: Experimental Scenario of Sally-Anne Task.
Source: Baron-Cohen et al. (1985)

(2) **Deceptive Box**
This task was introduced by Perner et al. (1987) and consisted mainly in showing the tested child a candy box of a well known brandy, which really contains matches. When the child discovers that there were not candies in there, the box is closed and then is asked to answer about his previous belief and how other child will think about the content of the chocolate box. Again, the idea of the task is to evaluate the ability of the child to put his self in someone else's mental beliefs. For this task, the age of passing it is 4 years and 3 months.

(3) **Pretense, Desire and Think**
Lillard and Flavell (1992) suggested these tasks in which the mental state of belief was replaced by others like pretend, want or desire, think, dream and looks like in a false-belief context to find specific conditions that can explain what elements makes it difficult for 3 year old children to succeed in the false belief tasks. Thus, the procedure involves telling the child that a person thinks that X is the case and learn that Y is the real situation. Then, they are asked about what the person thinks the object is. The same is applied for the other mental states. Also, regarding this task, only Pretense and Desire are known to be passed at 2 years and a half, whereas Think can be passed at 4 years old.

(4) **Narrative**
The task was proposed by Lewis et al. (1994) to evaluate the performance of children in false belief tasks based on their narrative background. According the authors, children can understand others' mental states once they have been able to pool picture frames into a narrative history. Thus, the task is presented as pictures in a book showing a similar procedure explained in the Location Change task. According to the author, the task can be passed at the age of 3 years and 9 months.

(5) **Nonverbal and Verbal**
Call and Tomasello (1999) suggested these two last tasks in order to avoid the difficulty that traditional false belief tasks (location change or deceptive box) involved like skills of inhibition or executive function in tested children. The authors believe that if the tasks do not account too much linguistic and inhibition abilities, younger children will perform more sensitively. The procedure can be seen as a new version of the Location Change false belief task which involves in both cases a communicator, a hider, a reward and two boxes. For the verbal test, the child is shown how the hider hides the reward in one box and then moves it when the communicator is outside the room. Conversely, for the nonverbal test, the hider only switches the boxes when the communicator is out of view. As in the traditional tasks, the question of interest relies on evaluate if the child understood that the communicator has other belief about the real location of the reward. This task can be passed at 4 years and 6 months.

In summary, the above tests are administered to children because they are simplified versions of the task. If children show developments in these tests over third year, this will make a clear contribution to the literature on Theory of Mind.

Chapter 3

Exploratory Analysis

This chapter gives a detailed description of the data and points out some important exploratory aspects such as the responses pattern, the preschooler performance over time, the correlation between the items across time and a mixed effects logistic regression to explore the evolution of the correct response of children.

3.1 Data Description

3.1.1 Participants

The data under study comprises the responses of 13 mental state questions made to 86 British children (Female = 41, Male = 45) in 2003 recruited from different preschools and day nurseries located in Northern Lancashire. Before conducting the study, consent was gained from children's parents or guardians. Moreover, the mental state tasks were given three times in intervals of 4 months. At first time, children were between ages of 30 to 33 months (Mean = 31.15, SD = 1.19) and belonged to 15 institutions. In the second time, children were aged between 34 and 37 months (Mean = 35.15, SD = 1.19). However, four children had moved to different nurseries, in which they were now tested after permission of their parents and the person in charge of the institutions. Therefore, 19 nurseries comprised the study in the second phase of testing. For the last time, children were from 38 to 41 months (Mean = 39.17, SD = 1.20). During this last phase, 2 children had changed nurseries and other children left the nursery, but they were tested in their new institutions and the last at his home after permission of their parents. Moreover, unfortunately two children were considered dropouts since they had left the nursery and moved away from the study area. It has also need to be mention that even if it was reported only two dropouts, there was actually 22 incomplete observations in the data base.

3.1.2 Measures

Each children was given 8 mental state tasks, which made up 13 questions of interest in total for each time, as specified below. For each correct response, the child received a score of '1' and a score of '0' if it was an incorrect response.

(a) **Standard Location Change:** For the first time, the child was presented with two different coloured boxes with lids (one pink and one blue), a pot of honey and two

well known toy characters, Tigger and Winnie the Pooh, all of them placed on the table. They were shown that Winnie the Pooh had a pot of honey and told that he was tired so he put his honey into the pink box and went to sleep. While he was gone, cheeky Tiger changed the honey from the pink box to the blue box in front of child view. On Winnie the Pooh's return, the child was asked with the following mental state question *"Where will Winnie the Pooh look for his honey first?"*. For the second and third time, the procedure was similar, but the toy characters, containers and the object were different. These were two opaque plastic containers with lids (one yellow and one red), a small ping-pong ball and the characters Jake and Doodle for Time 2. For Time 3, two opaque children's beakers with lids (one purple and one green), the characters Eeyore and Piglet and Eeyore's removable tail. In each case the first mentioned characters are the persons who left the scene. Thus, the test questions were *"Where will Doodle look for his ball first?"* and *"Where will Eeyore look for his tail first?"*, respectively.

(b) **Deceptive Box:** In Time 1, a Smarties tube containing pencils were shown to the children and asked what they thought was inside. After they realized that there were pencils inside, they were asked two test questions:

- Other false belief: *"What will X (another child at the nursery) think is in the box?"*
- Self belief: *"What did you think was in the box at the beginning?"*

For Time 2 and Time 3, the procedure was the same but with different objects, which were a Walker's crisp packet containing a baby's dummy and a small six pack egg box with a sock inside, respectively.

The following three tasks will be explained together since they had the same procedure:

(c) **Pretense, Desire and Think:** In Time 1, the experimenter used three boxes of different colours (pink, turquoise and silver), three paired of objects (horse & spoon, car & key, lego & raisins) and three characters. He presented each closed box containing the second object of each pair mentioned before and told the child three statements: 'Lucy is pretending there is a horse in the pink box', 'Dotty wants to be a car in the turquoise box' and 'Edward thinks there is some lego in the box'. Then, he made sure the child repeated the pretence, desire and think respectively by asking a control question. After this, the child opened the boxes to found there was a spoon, key and raisins inside of each box, so the experimenter asked the test questions corresponding to pretence, desire and think: *"What is Lucy pretending is in the box"?*, *"What does Dotty want there to be in the box?"* and *"What does Edward think is in the box?"*. It should be noticed that each task and question was made in different moments meaning three distinct mental tasks. Also, the boxes, the paired of objects and characters were counterbalanced across children and across mental states of pretence, desire and think. At Time 2, the materials were changed to three small different coloured tubes

(yellow, blue and orange), the paired of objects now were orange & tissues, rabbit & lollipop, stickers & ball and another three characters. At Time 3, the materials were three distinct mini cereal boxes (rice krispies, sugar puffs and coco pops), three pair of contents (cowboy & duck, plastic milk bottles & cat, pennies & a tiger) and another three characters. However, in both phases the procedure and questions remained the same.

(d) **Narrative:** The children were presented with a seven-paged story book containing a false belief problem. The story at time 1 and 2 was the same, but with different character (depending on what the child wanted to call her), as follows:

Page 1: Displays Sarah and her cat standing next to a television in the living room, situated between the bedroom and kitchen doors.

Page 2: Shows Sarah in the bedroom putting her cat into a basket.

Page 3: Shows Sarah back in the living room watching TV all the afternoon.

Page 4, 5 and 6: Show the cat quietly leaving by the bedroom window, getting into the kitchen window and going to sleep on a chair.

Page 7: Displays Sarah ready to leave the television and look for her cat.

In this moment, the experimenter asked the test question *"Now, which room will Sarah go into to get her cat?"*.

At Time 3, the story was about a little girl called Debbie and her pet mouse:

Page 1: Displays Debbie and her pet mouse.

Page 2: Shows a green box, a red box and Debbie holding some cheese. The experimenter told the child that the boxes belong to Debbie and she was putting the cheese into the red box.

Page 3: Shows the two boxes and Debbie putting her mouse into the green box. She tells the mouse to stay there sleeping and that he can eat the cheese at dinner.

Page 4: Shows Debbie playing in the garden.

Page 5 and 6: Displays the mouse getting out of the green box and going into the red box. Then, he eats the cheese. The experimenter told the child that Debbie does not see this.

Page 7: Shows Debbie approaching the two boxes to get her mouse after leaving the garden.

The experimenter then asked the child the test question *"Now, where do you think Debbie will look for her mouse?"*.

(e) **Non-Verbal:** The materials for this task were two boxes, one toy, a marker and a puppet communicator called Freddie. The experimenter sat opposite the child with a barrier separating them. He hide the toy in one box behind the barrier in Freddie sight. Then, the experimenter switched the position of the boxes while Freddie was out of the area. Few seconds later, Freddie came back and placed the marker where he thinks is the hidden toy. The child was asked then with the test question *"Where is the toy?"*. Four trials were done for this task each time.

(f) **Verbal:** In this task, the materials were the same as mentioned in the previous task. However, the procedure changed a little bit since the experimenter changed the toy to the other box instead of just stwiching the boxes and he told the child that this was to trick the puppet. All this happened when Freddie was out of sight and when he returned, the experimenter asked the child the test question *"Which box will Freddie put his marker on when he comes back?"*. The test consisted of two trials each time.

A more detailed description of the measures can be read in the Doctoral Thesis of Lunn (2006). The information provided in this section of the thesis is only a summary of it.

3.1.3 Coding Format

The label for the 13 test questions related to the eight mental state tasks evaluated in the study is shown in the following table:

Table 3.1: Labels of Items

	Test Question	Label
1.	Non Verbal False Belief - Trial 1	NVFBTR1
2.	Non Verbal False Belief - Trial 2	NVFBTR2
3.	Non Verbal False Belief - Trial 3	NVFBTR3
4.	Non Verbal False Belief - Trial 4	NVFBTR4
5.	Verbal False Belief - Trial 1	VFBTR1
6.	Verbal False Belief - Trial 2	VFBTR2
7.	Narrative	NARTESQ
8.	Deceptive Box - Other	DBOTHEQ
9.	Deceptive Box - Self	DBSELFQ
10.	Standard Location Change	STANTES
11.	Pretense	PRETENSE
12.	Desire	DESIRE
13.	Think	THINK

3.2 Non-Longitudinal Analysis

3.2.1 Response Patterns

The exploratory analysis can start by looking first the response patterns of the assessed children. There are only eight possibilities of responses since the study consists of questions that allowed a correct or incorrect response and this was done three times.

Fig.3.1 shows the number of children having certain pattern of response by each item evaluated. Moreover, the patterns are arranged according to the number of children who answered that way. It can be clearly seen that the dominant pattern in general is 000 meaning not correct response at any time. Also, this pattern is more highlighted for the items called Deceptive Box (DBOTHQ, DBSELFQ) and Standard Location Change (STANTES). The pattern having all right responses, 111, is in fourth place; however, for the items related to Pretense and Desire questions is at the top. Thus, these two can be considered easy questions. On the other hand, the patterns 001 and 011 are also important to mention because they can tell us that the child improved across time in his responses. We can see that these patterns have modest quantities for almost all items. The other patterns 010, 101 and 110 are the last of the list and this could be because children can guess and give random answers. It is important to mention that the numbers showing in Fig.3.1 are computed considering only complete observations; thus, the sum per column is not the same across all items.

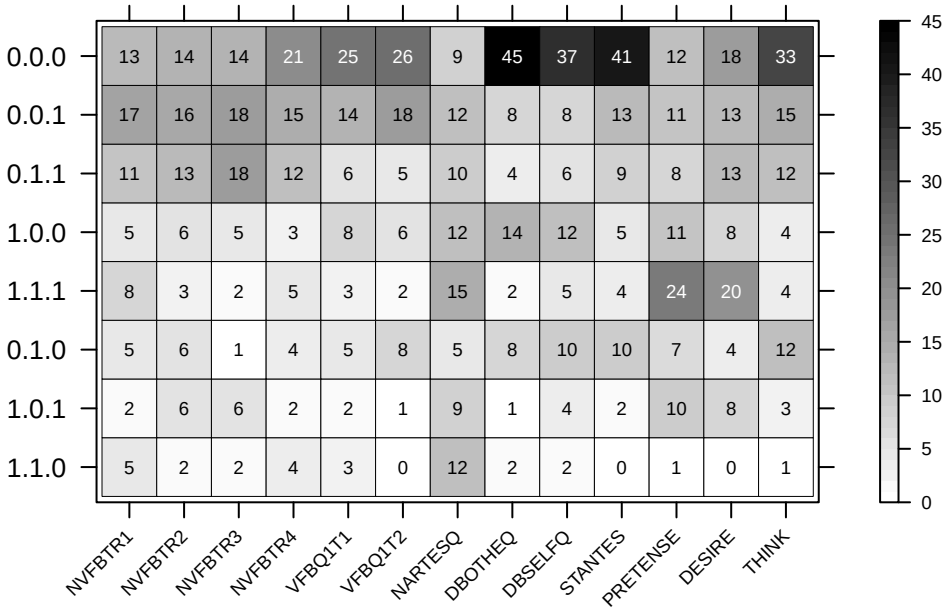

Figure 3.1: Response Patterns by Item

3.2.2 Total Performance

We analyse the data by inspecting the overall performance of all subjects as a function of time. This was computed for each child as an average of the Total Score obtained in all the questions [1] divided by the number of answered questions and we named it Success Index. The reason of the division was to not bias the performance across time since there was a quite amount of children that did not completed all the tasks. Thus, the success index for the i-th child is:

$$S_i = \frac{\sum_{j=1}^{n_i} y_{ij}}{n_i} \qquad i = 1, ..., 86 \quad j = 1, ..., n_i \tag{3.1}$$

where y_{ij} is the score obtained of individual i to item j and n_i is the number of items answered by individual i.

Fig.3.2 shows the performance for all the observed data including dropouts, incomplete profiles and for individuals that have all the measurements. It can be clearly noted that the overall mean performance increased over time. Moreover, the increase was steeper from Time 2 to Time 3 than from Time 1 to Time 2, which seemed reasonable since the children get older and they develop more the theory of mind.

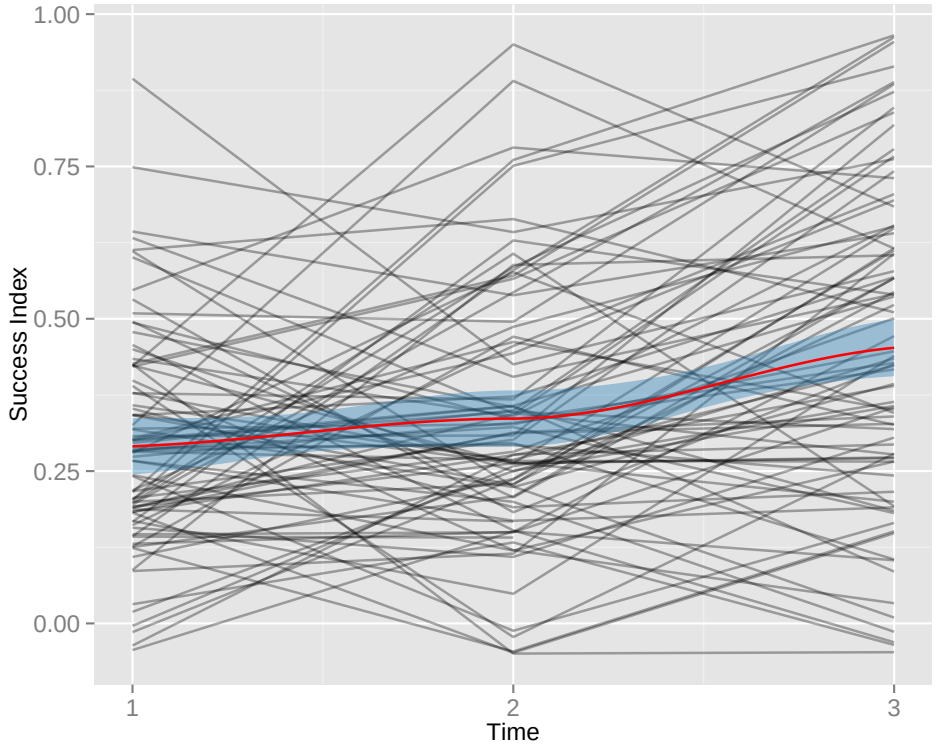

Figure 3.2: Total performance across time

[1] Questions and items are used interchangeable throughout the thesis.

3.2.3 Correlation Analysis

An inspection of the observed correlations trough time is needed to have an indication of possible latent factors and how they are associated at each time. Since all items are binary, polychoric correlations were computed as shown in Fig.3.3 and Fig.3.4. Large positive correlations are shades of blue while negative correlations are shades of red.

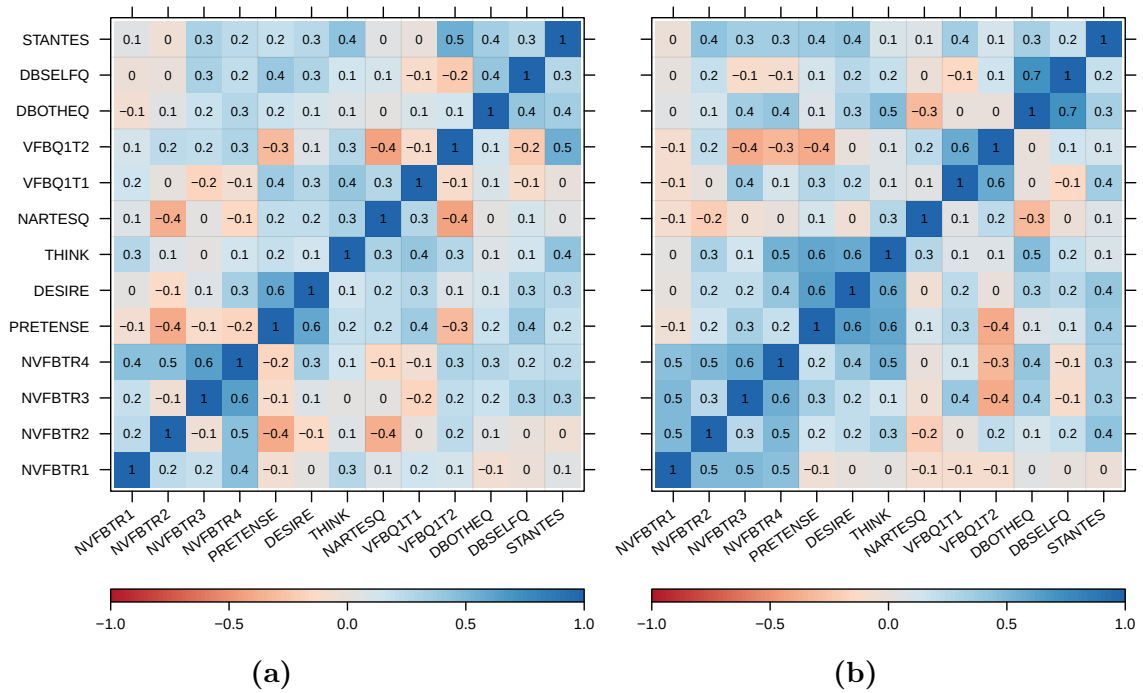

Figure 3.3: Polychoric correlations between items for (a) Time 1 and (b) Time 2.

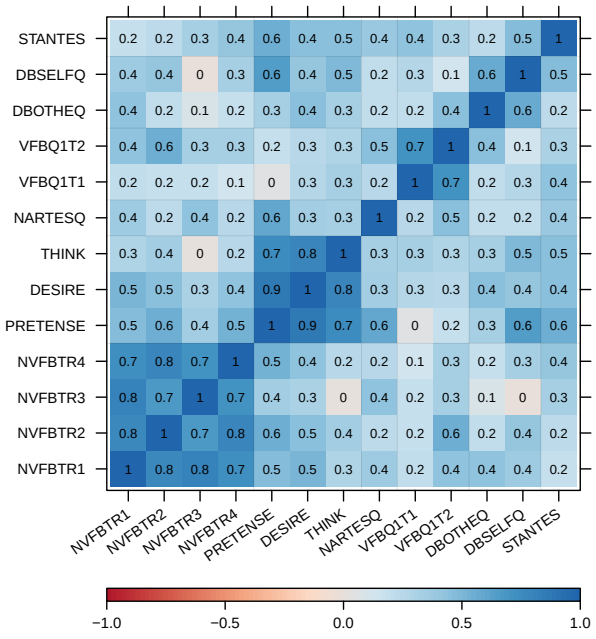

Figure 3.4: Polychoric correlations between items for Time 3.

It is clear from the plots that the correlation increases as time goes by and that the increment is more remarkable for certain items. For example, in Time 3 (Fig.3.4) the four trials of non verbal false belief task have strong association (above 0.6) as well as the two trials of verbal false belief, the two questions of deceptive box and the questions of pretense, desire and think. Therefore, a block structure can be seen which will suggest that the items can be grouped into separate latent domains. Also, this kind of association is not too much highlighted in the previous times, but it is still present.

3.3 Generalized Linear Mixed Model

Mixed effects logistic regression was applied for each item in order to investigate the evolution of its satisfactory response. Thereby, the observation of individual i at time j, defined as Y_{ij}, is assumed to come from a Bernoulli distribution with π_i, where π_i is related to a linear predictor η_{ij} by the link function $logit(\pi_{ij}) = \eta_{ij}$. The linear predictor has the following form:

$$\eta_{ij} = \beta_0 + t_{ij}\beta_1 + U_i \quad i = 1, ..., 86 \quad j = 1, 2, 3 \tag{3.2}$$

where β_0 and β_1 are the fixed parameters representing the intercept and slope, respectively. The parameter U_i is a random effect to allow heterogenity between individuals with $U_i \sim N(0, \sigma_1^2)$. There is only one covariate which is the time and this was centered to obtain uncorralated parameters estimate, so the interpretation can be easier.

Table 3.2 shows the point and interval estimates for the fixed gradient and intercept sorted by the estimation value of the gradient. The significance of each item is also presented in the last column of the table. These results were obtained using the **lme4** package in the R free software.

Table 3.2: Estimates of Fixed Parameters

Item	Gradient	SE	95% CI Lower	95% CI Upper	P-value
NVFBTR3	0.825	0.194	0.444	1.206	0.000
THINK	0.760	0.201	0.366	1.154	0.000
NVFBTR4	0.698	0.199	0.307	1.089	0.001
NVFBTR2	0.673	0.184	0.312	1.034	0.000
STANTES	0.648	0.212	0.233	1.064	0.002
DESIRE	0.630	0.195	0.247	1.013	0.001
NVFBTR1	0.618	0.186	0.252	0.984	0.001
VFBQ1T2	0.527	0.207	0.122	0.932	0.011
VFBQ1T1	0.309	0.186	-0.056	0.675	0.097
PRETENSE	0.209	0.174	-0.132	0.550	0.230
DBSELFQ	0.015	0.187	-0.351	0.380	0.938
NARTESQ	-0.025	0.157	-0.332	0.282	0.874
DBOTHEQ	-0.148	0.202	-0.543	0.247	0.464
Item	**Intercept**	**SE**	**95% CI Lower**	**95% CI Upper**	**P-value**
NVFBTR3	-0.416	0.158	-0.727	-0.106	0.009
THINK	-1.061	0.203	-1.458	-0.664	0.000
NVFBTR4	-0.653	0.189	-1.024	-0.283	0.001
NVFBTR2	-0.524	0.153	-0.823	-0.224	0.001
STANTES	-1.442	0.249	-1.931	-0.954	0.000
DESIRE	0.004	0.211	-0.409	0.417	0.986
NVFBTR1	-0.364	0.162	-0.682	-0.046	0.025
VFBQ1T2	-1.239	0.201	-1.632	-0.846	0.000
VFBQ1T1	-0.959	0.175	-1.302	-0.616	0.000
PRETENSE	0.239	0.184	-0.121	0.601	0.194
DBSELFQ	-1.191	0.213	-1.609	-0.773	0.000
NARTESQ	0.129	0.135	-0.135	0.393	0.339
DBOTHEQ	-1.574	0.233	-2.031	-1.117	0.000

As we can see, eight items had significant increase in success probability over time. Fig.3.5 displays an error bar plot of these results, also. It can be noticed that even though the third trial of non verbal false belief task (NVFBTR3) has an important increase, it

does not have a positive intercept. This means that by Time 2, the proportion of subjects answering correctly this task is below the fifty percent. It should be remembered that we analysed the intercept based on Time 2 because of the centering process of the covariate Time in the model. This pattern of behaviour also happens for the other trials of the task as well as for Think, Desire, second trial of Verbal (VFBQ1T2) and Standard Location Change (STANTES). However, it is remarkable that for STANTES the proportion is very low, although it has grown in success.

On the other hand, it is noted on the plot that despite the non significance increase on success of the tasks Pretense and Narrative (NARTESQ), the intercept is very high meaning that the proportion of children who had right responses was mildly high. A non significance growth can also be considered as a constant proportion over time. Nevertheless, this did not happen for the two test questions of Deceptive Box (DBSELFQ and DBOTHEQ). Conversely, both the gradient was not significant and the intercept was negative. This could mean that these two items are very difficult and the improvement on the score did not happen over time.

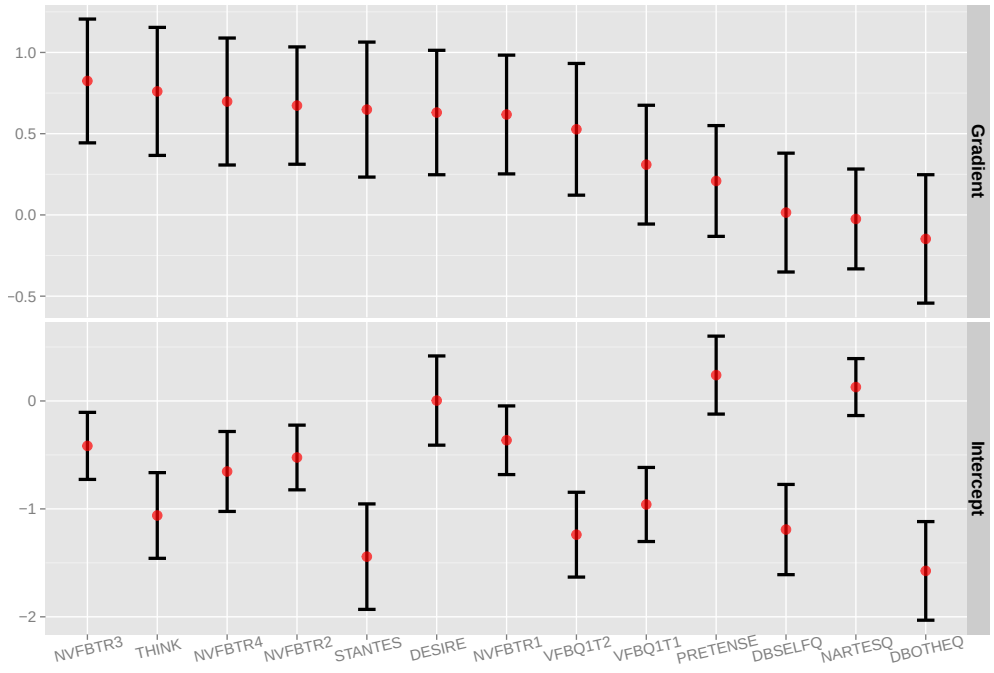

Figure 3.5: Point and Interval Estimate by Item in Logit Scale.

3.4 Conclusion

In this chapter, the main findings were that there is an increasing trend in the general performance of children, and that the items are correlated at each time point with a remarkable increase and block structure of the items in the third time.

Chapter 4

Multidimensional Item Response Modeling

In this chapter a brief description of Item Response Theory (IRT) is introduced and some models are explained concerning binary data. Then, the main focus relies on the extension to Multidimensional Item Response Theory (MIRT), specially for binary data. Finally, this methodology is applied to our objectives in the area of Theory of Mind. Specifically, to find the number of factors that characterise Theory of Mind by using the R package **mirt**.

4.1 Introduction to Item Response Theory

Item Response Theory is a set of latent variable techniques extensively used in educational and psychological areas. Its main concern is on measure a latent construct which is not observable and can only be measured indirectly through some manifest variables like a set of items (Fox, 2010). The basic idea of IRT is to analyse the interaction between the individual latent ability with the items characteristics (difficulty, guessing, etc.) by mapping a probabilistic model. There are three main assumptions concerning Item Response Theory:

1. The unidimensionality of the latent ability defined as θ that determines the item responses.

2. A change in the probability of a specific response because of a change in the latent variable is completely described by the Item Characteristic Curve (ICC). Hence, the ICC outlines how the probability of an item response varies regarding changes in the latent trait (Fox, 2010).

3. The Local Independence assumption stating that when the underlying latent trait is held constant, the responses to a pair of items are statistically independent. Moreover, when the unidimensionality assumption is true, local independence holds (Fox, 2010). Let $\mathbf{X}_i$ be a random vector, with observed values $\mathbf{x_i} = (x_{i1}, x_{i2}, ..., x_{in})$, of n item responses for the i-th individual with ability parameter θ_i. Then, this assumption can be stated as:

$$P(\mathbf{x_i} \mid \theta_i) = P(x_{i1} \mid \theta_i)P(x_{i2} \mid \theta_i) \ ... \ P(x_{in} \mid \theta_i) = \prod_{j=1}^{n} P(x_{ij} \mid \theta_i) \tag{4.1}$$

It is important to know how the data is obtained by expressing the likelihood function. Thereby, the probability of an individual i with observations $\mathbf{x_i} = (x_{i1}, x_{i2}, ..., x_{in})$ from a set of n items has the following form:

$$\begin{aligned} L_i(\mathbf{\Psi}, \theta_i) &= Pr(\mathbf{x_i} \mid \theta_i, \mathbf{\Psi}) \\ &= \prod_{j=1}^{n} Pr(x_{ij} \mid \theta_i, \Psi_j) \end{aligned} \tag{4.2}$$

Where Ψ_j is the vector of all item parameters for item j. Therefore, the probability of all the observed data is:

$$L(\mathbf{\Psi}, \theta) = \prod_{i=1}^{N} L_i(\mathbf{\Psi}, \theta_i) \tag{4.3}$$

Different IRT models have been proposed, but the focus of this thesis is only in the models corresponding to dichotomous data, which will be describe and enumerate in the next sub-section.

4.1.1 Binary Item Response Models

One-Parameter Logistic Model (1PL)

It is the simplest and most widely used item response model for this kind of data. In this model, the probability of a correct response for individual i with ability level θ_i is defined as:

$$P(x_{ij} = 1 \mid \theta_i, \alpha, d_j) = \frac{exp\{\alpha(\theta_i - d_j)\}}{1 + exp\{\alpha(\theta_i - d_j)\}} \tag{4.4}$$

Where d_j is the difficulty parameter for item j which describes how much ability an individual should have in order to have a 0.5 of probability to answer correctly such item and α is the discrimination or slope parameter expressing the relationship power between the latent ability and the item j. It has to be pointed out that for this model, the slope parameter remains constant for all the items and also the each latent ability θ_i is assumed to come from a standard normal distribution $N(0, 1)$. When α takes the fixed value of one and the ability is not treated as a random variable, the resulting model is called Rasch. Therefore, the interpretation for the d_j in the Rasch Model is that as higher its value, the easier the item j (Fox, 2010).

Two-Parameter Logistic Model (2PL)

This model is a generalization of the previous model and allows the discrimination parameter to vary from item to item. Then, the probability of an individual i to answer

correctly the item j is:

$$P(x_{ij} = 1 \mid \theta_i, \alpha_j, d_j) = \frac{1}{1 + exp\{-D\alpha_j(\theta_i - d_j)\}} \tag{4.5}$$

Where D is a scaling constant with a common value of 1.7 and is considered a historical artefact used to have the same scale from the logistic model to the normal ogive model (Wirth and Edwards, 2007). This model holds an important fact that the higher (lower) the discrimination parameter, the (less) better the item is able to distinguish between low and high ability levels (Fox, 2010).

Fig. 4.1 displays the Item Characteristic Curves corresponding to the equations 4.4 and 4.5 considering two items with specific parameters. As it can be noticed, as more difficult the item, the ICC tends to be more placed to the right. Moreover, the ICC curves for the two items cross in the 2PL model because of the mismatch on the discrimination parameter.

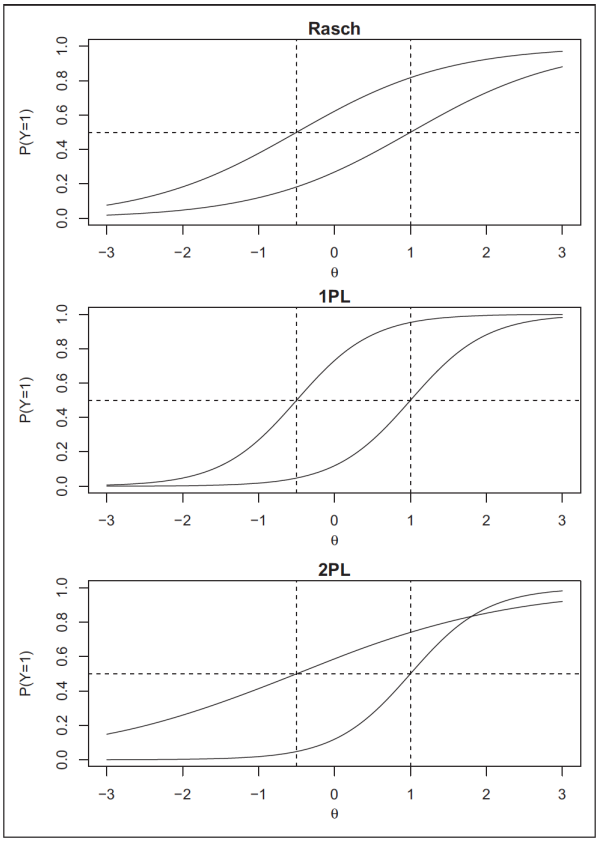

Figure 4.1: Item Characteristic Curves (ICC) for three models with two items having each one difficulties of $d_1 = -0.5$ and $d_2 = 1$. The discrimination in the Rasch model is $\alpha = 1$ and $\alpha = 2$ in the 1PL for both items. For the 2PL, this parameter is $\alpha_1 = 0.7$ and $\alpha_2 = 2$.
Source: Titman et al. (2013)

Three-Parameter Logistic Model (3PL)

This model is an extension of the 2PL, which incorporates a guessing parameter η_j to the probability of success. This is specially for multiple choice tests since the individuals can choose an answer just by chance (Curtis, 2010). As a result, the ICC moves up a little bit in the vertical axis representing the probability to have a right answer to item j having a low ability. The model is then defined as follows:

$$P(x_{ij} = 1 \mid \theta_i, \alpha_j, d_j, \eta_j) = \eta_j + (1 - \eta_j)\frac{1}{1 + exp\{-\alpha_j(\theta_i - d_j)\}} \tag{4.6}$$

For this thesis, the model considered is the 2PL but in the multidimensional approach, which will be explained in the sub-section 4.2.1.

4.1.2 Estimation Procedure

IRT models can be estimated using four basic techniques: Joint Maximum Likelihood (JML), Conditional Maximum Likelihood (CML), Marginal Maximum Likelihood (MML), and Bayesian estimation with Markov Chain Monte Carlo. Basically, all these methods rely strongly on the independence between individuals and on the local independence assumption stated before (Johnson, 2007).

In general terms, the first two techniques consider both the latent ability and the item parameters as unknown fixed parameters. The JML consists in estimating simultaneously the item parameters and the subject's abilities by an iterative procedure. Moreover, this technique adds some constraints to the model parameters to overcome the drawback of a non-identifiable model (not unique solution in the maximization) (Johnson, 2007). The results obtained in the estimation, however, are inconsistent. Even though the sample of examinees increases, the estimates will remain biased (Andersen, 1970; Ghosh, 1995).

On the other hand, the CML is an alternative method suggested by Andersen (1970). His methodology lies on simplifying the likelihood by conditioning on a sufficient statistic for the underlying subject's ability in the sample. Nonetheless, despite the consistent estimates obtained, the procedure gets more difficult in complex models like the two parameter logistic since it is not easy to find a simple sufficient statistics in there (Johnson, 2007).

The Marginal Maximum Likelihood is commonly used and the idea is to remove the ability from the likelihood function, expressed in Eq. 4.3, by integrating it out. To be able to do this, the latent ability θ has to be considered as a random variable, usually assumed to have a $N(0, 1)$ distribution. Therefore, the individual likelihood stated in Eq. 4.2 can be unfold to get the marginal probability of observing the item response vector

$\mathbf{x_i} = (x_{i1}, x_{i2}, ..., x_{in})$ as follows:

$$\begin{aligned} L_i(\mathbf{\Psi}) &= Pr(\mathbf{x_i} \mid \mathbf{\Psi}) \\ &= \int_\Theta Pr(\mathbf{x_i}, \theta_i \mid \mathbf{\Psi}) d(\theta_i) = \int_\Theta Pr(\mathbf{x_i}, \theta \mid \mathbf{\Psi}) d(\theta) \qquad (4.7) \\ &= \int_\Theta Pr(\mathbf{x_i} \mid \theta, \mathbf{\Psi}) Pr(\theta \mid \mathbf{\Psi}) d(\theta) \\ &= \int_\Theta Pr(\mathbf{x_i} \mid \theta, \mathbf{\Psi}) Pr(\theta) d(\theta) \qquad (4.8) \end{aligned}$$

Then, the marginal likelihood of the item parameter vector $\mathbf{\Psi}$ can be computed by taking the product of Eq. 4.8 over all individuals as:

$$L(\mathbf{\Psi}) = \prod_{i=1}^{N} L_i(\mathbf{\Psi}) = \prod_{i=1}^{N} \int_\Theta Pr(\mathbf{x_i} \mid \theta, \mathbf{\Psi}) Pr(\theta) d(\theta) \qquad (4.9)$$

The MML estimates are therefore obtained by maximising the previous likelihood with respect to the item parameters $\mathbf{\Psi}$. Unfortunately, the integral in this expression can not be solved analytically, so numerical integration techniques are required. Thus, Gauss-Hermite quadrature is employed (Bock and Aitkin, 1981) obtaining the expression:

$$\int_\Theta Pr(\mathbf{x_i} \mid \theta, \mathbf{\Psi}) Pr(\theta) d(\theta) \approx \sum_{q=1}^{Q} Pr(\mathbf{x_i} \mid \mathbf{\Psi}, K) g(K_q) = \tilde{P}_l \qquad (4.10)$$

Where K_q are the nodes and $g(K_q)$ are the respective weights. Then, to continue with the estimation and considering only the case of binary data, let define u to be the unique response patterns found in the sample and r_u the number of subjects who respond the pattern r_u. As a result, the observed likelihood expressed in Eq. 4.9 can be arranged now to the form:

$$L(\mathbf{\Psi} \mid \mathbf{X}) = \frac{N!}{r_1! r_2! ... r_u!} \tilde{P}_1^{r_1} \tilde{P}_2^{r_2} \ ... \ \tilde{P}_u^{r_u} \qquad (4.11)$$

An EM algorithm is applied to find out the item parameter estimates with the corresponding algorithm:

1. Initialize with specific values of the item parameters.
2. **E-Step:** The expected values of the reponse patterns are computed for each item, conditioned on the current estimates and the data.
3. **M-Step:** Maximise the log-likelihood using the information of the E-Step.

Repeat the above procedure until convergence is attained. For more details related to EM algorithm look the paper of Bock and Aitkin (1981).

Finally, the Bayesian approach considers both the latent ability and the item parameters as randoms. Hence, the item parameters $\mathbf{\Psi}$ have prior distributions that emphasize the uncertainty about their true value before observing the data. Once the sample likelihood is computed, the prior distributions are updated and the posterior distributions can be obtained. Assuming independence between the prior distributions, the joint posterior density of the parameters of interest after applying the Bayes' Theorem is:

$$Pr(\theta, \mathbf{\Psi} \mid x) = \frac{Pr(x \mid \theta, \mathbf{\Psi})\pi(\theta, \mathbf{\Psi})}{Pr(x)} \tag{4.12}$$

$$= \frac{Pr(x \mid \theta, \mathbf{\Psi})\pi(\theta)\pi(\mathbf{\Psi})}{Pr(x)} \tag{4.13}$$

The computation of the above distribution results analytically intractable and therefore the marginal posterior distributions have to be taken into account by marginalising out the not wanted parameter in the posterior distribution.

$$Pr(\theta \mid x) = \int \pi(\theta, \mathbf{\Psi} \mid x) d\mathbf{\Psi} \tag{4.14}$$

$$Pr(\mathbf{\Psi} \mid x) = \int \pi(\theta, \mathbf{\Psi} \mid x) d\theta \tag{4.15}$$

Nonetheless, summarizing the marginal posterior is difficult since the mathematical forms are not known. As a result, MCMC methods like Metropolis-Hastings have to be employed to approximate the posterior distributions. Review the book of Fox (2010) for more details on the procedure of the MCMC algorithm.

In the 2PL model, normal priors are common assumed for the discrimination and difficulty parameters. Then, for j=1,2,...,n items, we have:

$$\alpha_j \sim N(\mu_\alpha, \sigma_\alpha^2) I(\alpha_j > 0)$$

$$d_j \sim N(\mu_d, \sigma_d^2)$$

4.2 Multidimensional Item Response Theory

Multidimensional Item Response Theory (MIRT) can be considered as an extension of Item Response Theory where the latent trait is now treated as a vector of latent constructs. This was done because some tests require several abilities to answer correctly (Fox, 2010) and of course, because of the clearly multidimensional nature of many psychological constructs (Chalmers, 2012).

4.2.1 MIRT Model for Binary Data

The model defined in Eq. 4.5 can be generalised to a multidimensional approach. Let $i = 1, ..., N$ participants, $j = 1, ..., n$ test items, m latent factors $\boldsymbol{\theta}_i = (\theta_{i1}, ..., \theta_{im})$ with associated item slopes $\boldsymbol{\alpha}_j = (\alpha_1, ..., \alpha_m)$, d_j the item intercept and D a scaling adjustment (usually 1.702) (Reckase, 2009), then the probability of an individual i answering correctly the binary item j can be stated as:

$$\Phi(x_{ij} = 1|\boldsymbol{\theta}_i, \boldsymbol{\alpha}_j, d_j) = \frac{1}{1 + exp[-D(\boldsymbol{\alpha}_j^T \boldsymbol{\theta}_i + d_j)]} \tag{4.16}$$

In order to define the likelihood, all the equations defined for the unidimensional IRT will be extend here. Hence, letting again $\boldsymbol{\Psi}$ be the set of all item parameters, the conditional distribution of the i-th response pattern vector, $\mathbf{x_i} = (x_{i1}, x_{i2}, ..., x_{in})$, can be computed similar to Eq. 4.2:

$$\begin{aligned} L_\ell(\boldsymbol{\Psi}, \boldsymbol{\theta}) &= Pr(\mathbf{x_i} \mid \boldsymbol{\theta}, \boldsymbol{\Psi}) \\ &= \prod_{j=1}^{n} Pr(x_{ij} \mid \boldsymbol{\theta}_i, \boldsymbol{\Psi}) \end{aligned} \tag{4.17}$$

From the above expression, the marginal distribution can be derived by integrating out the m latent ability as:

$$\begin{aligned} L_\ell(\boldsymbol{\Psi}) &= \int_\Theta L_\ell(\boldsymbol{\Psi}, \boldsymbol{\theta}) Pr(\boldsymbol{\theta} \mid \boldsymbol{\Psi}) d(\boldsymbol{\theta}) \\ &= \int_{-\infty}^{\infty} ... \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} Pr(\mathbf{x_i} \mid \boldsymbol{\theta}, \boldsymbol{\Psi}) Pr(\boldsymbol{\theta}) d(\boldsymbol{\theta}) \end{aligned} \tag{4.18}$$

Then, the likelihood function for the observed data $\mathbf{X}$, a $N \times n$ matrix, can be computed by taking the product of the m-fold integrals in Eq. 4.18 over all individuals:

$$L(\boldsymbol{\Psi}; \mathbf{X}) = \prod_{i=1}^{N} L_\ell(\boldsymbol{\Psi}) = \prod_{i=1}^{N} \int_{-\infty}^{\infty} ... \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} Pr(\mathbf{x_i} \mid \boldsymbol{\theta}, \boldsymbol{\Psi}) Pr(\boldsymbol{\theta}) d(\boldsymbol{\theta}) \tag{4.19}$$

Eq. 4.19 is thus used to find the marginal likelihood estimates.

4.2.2 Parameter Estimation

The parameter estimation for the MIRT model is of the same way as explained in the marginal maximum technique for the unidimensional IRT model using the Expectation-Maximisation algorithm. Therefore, we need first to approximate the m-fold integrals expressed in Eq. 4.18 for unique response patterns employing a m-fold Gauss-Hermite quadrature. The expression in Eq. 4.10 for unidimensional IRT takes now the following

form:

$$\tilde{P}_\ell = \sum_{qm=1}^{Q} ... \sum_{q2}^{Q} \sum_{q1}^{Q} Pr(\mathbf{x_i} \mid \mathbf{\Psi}, \mathbf{K}) g(K_{q1}) g(K_{q2}) ... g(K_{qm}) \tag{4.20}$$

Where $K_{q1}, K_{q2}, ..., K_{qm}$ are the nodes and $g(K_{q1}), g(K_{q2}), ..., g(K_{qm})$ are their respective weights. Then, the likelihood function for the observed data can be arranged based on u unique response patterns with r_u number of subjects getting the specific pattern r_u as follows:

$$L(\mathbf{\Psi} \mid \mathbf{X}) = \frac{N!}{r_1! r_2! ... r_u!} \tilde{P}_1^{r_1} \tilde{P}_2^{r_2} \ ... \ \tilde{P}_u^{r_u} \tag{4.21}$$

Having the corresponding observed data likelihood, the item parameters can be found by differentiating in ψ_j and integrating out the m latent factors of $\boldsymbol{\theta}$, which will lead to the EM algorithm steps (the details for this are explained in Chalmers (2012)).

It has to be mentioned that the EM algorithm is only useful when there are not too much dimensions present in the latent ability $\boldsymbol{\theta}$. Titman et al. (2013) point out that 20 quadrature nodes may be necessary to approximate the integral with a reasonable accuracy in the unidimensional IRT. However, as the number of dimensions increase, so does the number of nodes in an exponentially way (Chalmers, 2012). For example, if the latent trait has m dimensions, then the number of nodes would be 20^m leading to high computation times.

In order to overcome this drawback, the use of Metropolis-Hastings Robbins-Monro (MH-RM) algorithm was proposed to estimate the item parameters for both the Exploratory (Cai, 2010b) and Confirmatory Factor Analysis (Cai, 2010a).

4.2.3 MIRT as Item Factor Analysis

The Multidimensional Item Response Models can be considered as extensions of linear factor analysis. Therefore, it is reasonable for these to have an Exploratory and Confirmatory phase.

In the Exploratory phase, the number of dimensions are not known and they will be estimated by comparing nested models or by rotating the factor loading matrix to get a more remarkable structure (Bock et al., 1988).

On the other hand, in the Confirmatory phase, there is an intuition that more than one dimension is presented in the set of items. According to Adams et al. (1997), it can be distinguished two types of Confirmatory Item Factor Analysis:

- Between-Item Multidimensionality, the items belong to only one latent dimension, but these can correlate between each other. See Fig. 4.2 for a diagram description.
- Within-Item Multidimensionality, the items measure more than one latent factor. A well known model that relies on this category is the Bifactor Model in which the items belong to a general latent variable, but also the items are grouped together in independent clusters as seen in Fig. 4.3. This model was explained by Gibbons and Hedeker (1992) for binary data considering the Full Information approach.

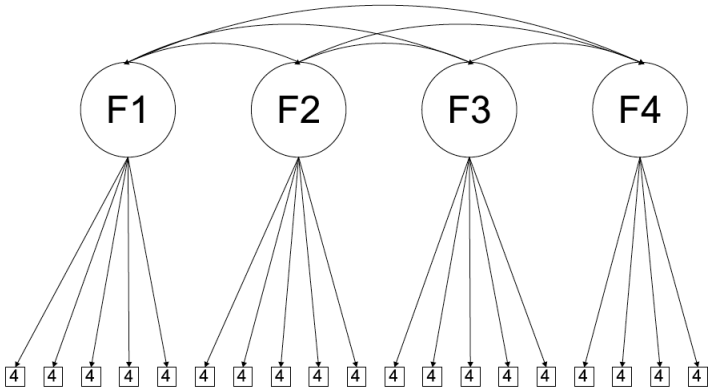

Figure 4.2: Within-Item model path for 20 items with 4 response categories each. **Source:** Edwards (2010)

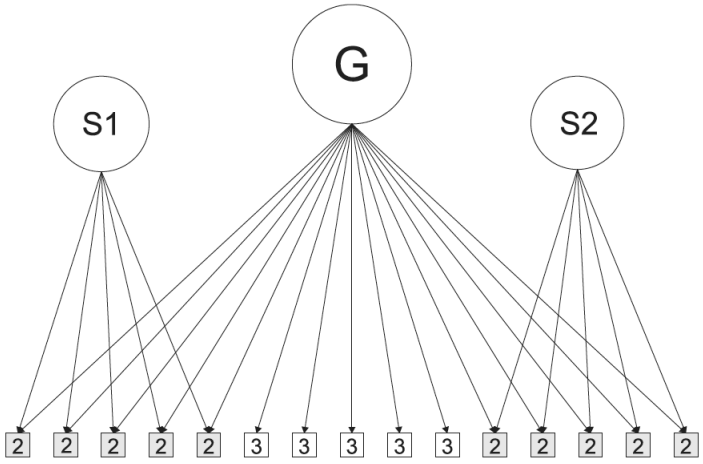

Figure 4.3: Bifactor model path for 15 items where the number indicates the responses categories in each item. **Source:** Edwards (2010)

4.3 Application to Theory of Mind

In this section, all the theory explained before is applied to our set of 13 binary items. First, an exploratory factor analysis is done to look for the number of latent dimensions in what Theory of Mind can be divided. Then, confirmatory factor analysis proceeds considering a Bifactor model. All the procedure was done in the free software R (version 3.2.1, 2015-06-18, "World-Famous Astronaut") using the **mirt** package version 1.10 (Chalmers, 2012). For this part, all the data was considered including the incomplete profiles since the package used take care of this internally.

4.3.1 Exploratory Factor Analysis

Initially, there is not previous knowledge of the number of dimensions that comprises Theory of Mind. Therefore, three factor models considering 2, 3 and 4 dimensions were done to find the appropriate final model with sensible latent dimensions. However, when the number of dimensions considered in the modelling is high, the author Chalmers (2012) recommends to estimate the parameters using the Metropolis-Hastings Robbins-Monro algorithm instead of the traditional EM approach. Thus, in order to have comparable results, this estimation approach was applied to all the different factor models.

Table 4.1 shows a nested comparison of the 4 models. The comparison was done taking into account five different criteria:

- *Akaike Information Criterion*, defined as $\text{AIC} = -2\log(L) + 2p$ including a penalization of two for every parameter.
- *Second Order Information Criterion*, it is based on the AIC, but with an additional correction for the sample size. The change in the formula depends on the complex of the fitted model. In case of a univariate linear model with normal residuals, it is computed as $\text{AICc} = \text{AIC} + \frac{2p(p+1)}{n-p-1}$.
- *Bayesian Information Criterion*, similar to the AIC but with a stronger penalization on the number of parameter, the $\text{BIC} = -2\log(L) + p\log(n)$.
- *Sample Size Adjusted Bayesian Information Criterion*, a sample-size adjusted version of the BIC defined as $\text{SABIC} = -2\log(L) + p\log(\frac{n+2}{24})$
- *Log-Likelihood Ratio Test*, to evaluate nested models as $\text{LRT} = 2\log(\frac{L_1}{L_0})$.

Considering the first four criteria just mentioned, the preferred model is the one with lowest value. Regarding the LRT, the test has to be significant to decide for the complex model.

Returning to the analysis of the Table 4.1, it can be pointed out that for a model with three factors the AIC, AICc and SABIC decrease in comparison to the model with

2 factors. However, if we add one extra factor, only the AIC diminishes. On the other hand, the LRT resulted significant at $\alpha = 0.05$ in both situations, comparing the 3 factor model to the 2 factor model and the 4 factor model to the 3 factor model. These results suggest that a 3 factor model will be reasonable, but we opted for the 4 factor model since according to the structure displayed, it becomes more sensible in relation to the psychological context.

Table 4.1: Nested Model Comparison

Model	AIC	AICc	SABIC	BIC	logLik	X^2	df	p
2 Factors	3667.51	3681.05	3682.05	3802.53	-1795.76			
3 Factors	3656.66	3680.21	3675.41	3830.75	-1779.33	32.86	11	0.00
4 Factors	3655.07	3690.83	3677.65	3864.70	-1768.54	21.58	10	0.02

Moreover, the percentage of explained variance is presented for the three models in Table 4.2. It is clear that as the number of latent dimensions incorporated increases, the explained variance also goes up. The increment, however, is a little bit more meaningful from the first model to the second than from this last to the model with three factors.

Table 4.2: Percentage of explained variance

Factors	2	3	4
% Variance	41.1	50	56.7

Table 4.3 shows the factor loadings for the chosen model with 4 dimensions after applying a varimax rotation. This rotation allows the factors to be orthogonal (meaning no correlation), which makes loads interpretation easier. Comparing the relatives loadings of each item in each of the four factors, it can be seen that factor F1 can be characterised by the first four items (NVFBTR1, NVFBTR2, NVFBTR3, NVFBTR4) as it is where the high loads are.

The same interpretation can be done for the remaining factors. Then, F2 is identified by the next three items which are PRETENSE, DESIRE and THINK. F3 is represented by items VFBQ1T1 and VFBQ1T2, and F4 is described by the two items related to Deceptive Box task (DBOTHEQ and DBSELFQ). However, the items of the Standard Location Change task (STANTES) and the Narrative Test (NARTESQ) does not belong to any particular factor since these do not hold a high load (above 0.5).

Now, that we know that the construct of Theory of Mind is comprised of four latent dimensions, it makes sense to continue the analysis with a confirmatory approach to estimate the values of interest, which is explaining below.

Table 4.3: Factor Loading Matrix after Varimax Rotation

Item	F1	F2	F3	F4
1. NVFBTR1	**0.72**	-0.10	-0.12	0.04
2. NVFBTR2	**0.67**	-0.08	-0.28	-0.23
3. NVFBTR3	**0.75**	-0.19	-0.07	0.20
4. NVFBTR4	**0.85**	-0.18	-0.09	-0.12
5. PRETENSE	0.13	**-0.94**	0.09	-0.05
6. DESIRE	0.27	**-0.73**	-0.15	-0.14
7. THINK	0.18	**-0.58**	-0.22	-0.19
8. NARTESQ	0.02	-0.32	-0.14	0.13
9. VFBQ1T1	0.05	-0.33	**-0.56**	0.27
10. VFBQ1T2	0.15	0.05	**-0.97**	-0.07
11. DBOTHEQ	0.13	-0.31	-0.07	**-0.49**
12. DBSELFQ	0.00	-0.39	-0.10	**-0.57**
13. STANTES	0.36	-0.28	-0.24	-0.23

4.3.2 Confirmatory Factor Analysis

A Bifactor model was applied as a confirmatory approach considering a model with a General dimension which involves all the items and four independent dimensions as it was finding in the previous subsection. The following table presents the factor loadings to this model.

Table 4.4: Factor Loading Matrix - Bifactor Method

Item	G	F1	F2	F3	F4
1. NVFBTR1	0.31	0.74			
2. NVFBTR2	0.42	0.59			
3. NVFBTR3	0.43	0.61			
4. NVFBTR4	0.55	0.66			
5. PRETENSE	0.53		0.63		
6. DESIRE	0.65		0.61		
7. THINK	0.57		0.36		
8. NARTESQ	0.28				
9. VFBQ1T1	0.41			0.56	
10. VFBQ1T2	0.37			0.65	
11. DBOTHEQ	0.37				0.88
12. DBSELFQ	0.39				0.47
13. STANTES	0.71				

Bifactor loadings for the four independent factors were practically similar to dominants loadings in the exploratory model with 4 factors (see Table 4.3). Even though, the items NARTESQ and STANTES do not belong to any factor, they have a weak (0.28)

and a very strong (0.71) relation with the latent construct of Theory of Mind, respectively.

A graphical way to show the dependencies of this type of confirmatory model is by representing it with a path plot as it is shown in Fig. 4.4. It gives a clearer view of how the items are grouped together in different factors and also how they are related to the general ability.

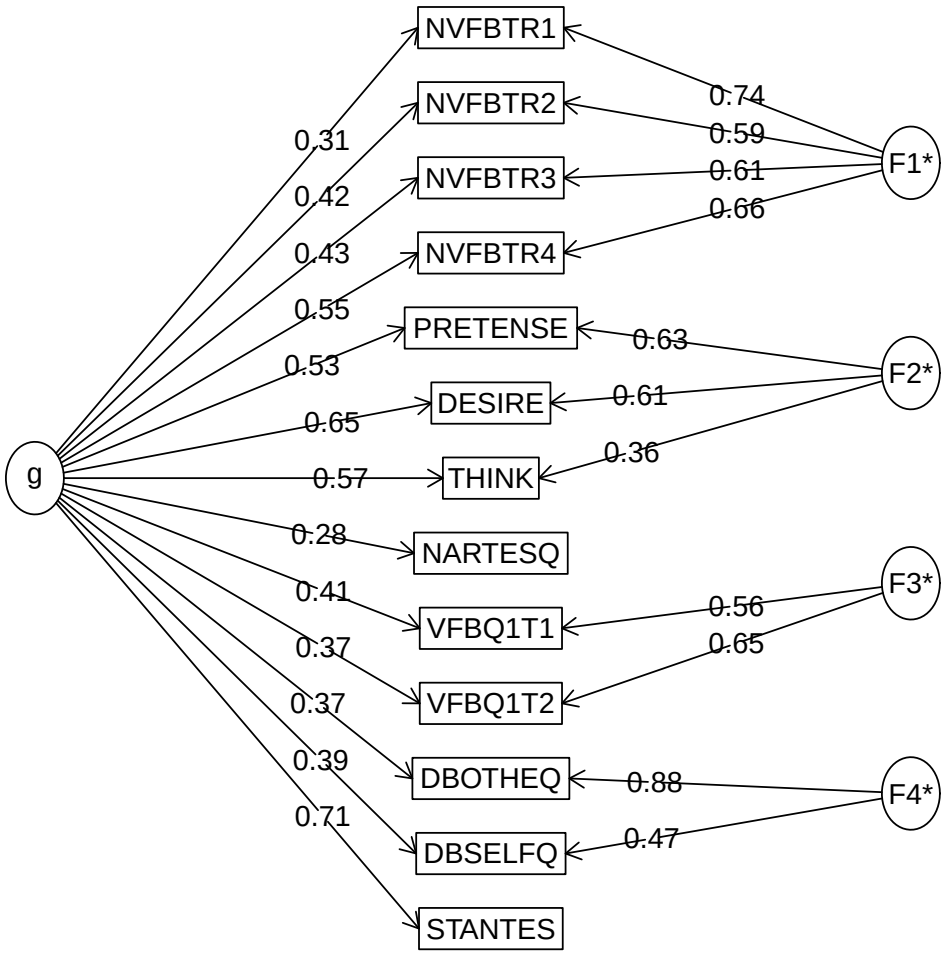

Figure 4.4: Bifactor model Path.
Use of **psych** package (Revelle, 2015) to do the structure, but not the output numbers.

4.4 Conclusion

This chapter drew out the dimensional reduction of Theory of Mind general ability. It was found out that it is comprised by 6 latent dimensions according to the exploratory and confirmatory factor analysis. Four main factors grouping 11 items (see Fig. 4.4) and two others containing one item each.

Chapter 5

The Two Stage Approach to determine Causality

In this chapter, we extend the one dimensional latent ability parameter of Theory of Mind to the six dimensions we have identified in the last chapter. In our first stage, we estimate the mean of this 6 dimensional vector latent ability factor. In the second stage, we regress the latent ability factors of times $t = 2, 3$ against the latent ability of the previous instant of time.

5.1 First Stage: Bayesian Longitudinal Model

We have seen so far that, in the context of Item Response Modeling, we obtain a unique single vector of responses for each examinee and all the items measure the same unidimensional ability. However, sometimes the objective is to analyse the evolution of certain ability in the examinees and therefore the same questionnaire is administered on multiple time points.

Regarding our data, recall that in the previous chapter, we found 4 dimensions that comprises the latent general ability of Theory of Mind and two free items that did not belong to any of the dimensions, but that they still explained the global ability. For the analysis in this chapter, we will consider those 2 free items as two additional factors. Therefore, 6 factors will be employed as latent unidimensional abilities of Theory of Mind.

5.1.1 The Likelihood

Let the ability vector in the factor f of T times be $\boldsymbol{\theta}_{i,f,1:T} = (\theta_{if1}, \theta_{if2}, \ldots, \theta_{ifT})'$, where θ_{ift} represents the ability of the i-th subject in the latent dimension f at time t, and the binary unique responses matrix $\boldsymbol{X}_{if}$ to n items for the same individual and dimension f defining as follows:

$$\begin{aligned}\boldsymbol{X}_{if} &= \begin{pmatrix}\boldsymbol{X}_{if1} & \boldsymbol{X}_{if2} & \boldsymbol{X}_{if3} & \ldots & \boldsymbol{X}_{ifT}\end{pmatrix} \\ &= \begin{pmatrix} x_{i11} & x_{i12} & x_{i13} & \ldots & x_{i1T} \\ x_{i21} & x_{i22} & x_{i23} & \ldots & x_{i2T} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x_{in1} & x_{in2} & x_{in3} & \ldots & x_{inT} \end{pmatrix}\end{aligned}$$

For example, in our data, the responses for subject i to the first dimension corresponding to Non Verbal False Belief tasks has the form:

$$\begin{aligned}\boldsymbol{X}_{if_1} = & \ (\boldsymbol{X}_{if_1 1} \quad \boldsymbol{X}_{if_1 2} \quad \boldsymbol{X}_{if_1 3}) \\ = & \begin{pmatrix} x_{i11} & x_{i12} & x_{i13} \\ x_{i21} & x_{i22} & x_{i23} \\ x_{i31} & x_{i32} & x_{i33} \\ x_{i41} & x_{i42} & x_{i43} \end{pmatrix}\end{aligned}$$

where the number of rows represents the number of items in the first dimension ($n_{f_1} = 4$) and the columns stand for each time point $t = 1, 2, 3$. The remaining factors have the same structure but different numbers of items ($n_{f_2} = 3, n_{f_3} = 2, n_{f_4} = 2, n_{f_5} = 1, n_{f_6} = 1$).

The response x_{ijt} is a random variable that comes from a Bernoulli distribution with probability p_{ijt} satisfying $\text{logit}(p_{ijt}) = \alpha_j(\theta_{i,f_j,t} - d_j)$, in where $i = 1, \ldots, N$ individuals, $j = 1, \ldots, n$ items and $t = 1, \ldots, T$ time points. Thus, the complete likelihood for our data can be computed as:

$$\begin{aligned} L(\boldsymbol{\theta}, \boldsymbol{\Psi}) &= P(X \mid \boldsymbol{\theta}, \boldsymbol{\Psi}) P(\boldsymbol{\theta}) \\ &= \prod_{i=1}^{N} \prod_{j=1}^{n} \prod_{t=1}^{T} p_{ijt}^{x_{ijt}} (1 - p_{ijt})^{1 - x_{ijt}} \times \prod_{i=1}^{N} P(\theta_{i,1:6,1:3}) \end{aligned} \tag{5.1}$$

There are three ways in which the second part of the likelihood expressed in Eq.5.1 can be parametrised, which I will briefly describe them in the next paragraphs.

Our first approach is to assume that the subjects' abilities for each factor f come from a multivariate normal distribution $\boldsymbol{\theta}_{i,f,1:T} \sim N(\boldsymbol{\mu}_\theta, \boldsymbol{\Sigma}_\theta)$ with a specific covariance structure. This can be the Homogeneous First-Order Autoregressive (AR(1)) which basically assumes a constant variance of the latent ability across time, but with an exponential correlation decrease as the lag between times increases. $\boldsymbol{\Sigma}_\theta$ is then defined for each dimension f in the 3 time points as,

$$\begin{aligned} \boldsymbol{\Sigma}_\theta &= \sigma_f^2 \boldsymbol{R}_f \\ &= (\sigma^2 \rho_f^{|t_1 - t_2|}) \ \text{ where } t_1, t_2 \in \{1, 2, 3\} \\ &= \sigma^2 \begin{pmatrix} 1 & \rho_f & \rho_f^2 \\ \rho_f & 1 & \rho_f \\ \rho_f^2 & \rho_f & 1 \end{pmatrix} \end{aligned}$$

Nonetheless, a problem of identifiability (not unique solution) is found in the model process, so it results necessary to add a restriction in the specification of the mean and variance of the latent ability at some particular time (Tavares and Andrade, 2006). According to Tavares and Andrade (2006), one commonly possibility is to establish $\boldsymbol{\mu}_\theta^{(1)} = 0$ and $\sigma^2 = 1$.

A second alternative is to use an Unstructured Covariance (no patterns) because researchers may occasionally be unwilling to specify a explicit covariance structure $\boldsymbol{\Sigma}_\theta$ for the latent abilities (Curtis, 2010). According to the literature, unstructured covariances commonly use Wishart distributions as priors, but in the context of Item Response Theory this leads to an identifiability problem. This problem can be overcome by using the Cholesky decomposition to the covariance matrix $\boldsymbol{\Sigma}_\theta = \boldsymbol{L}_\theta \boldsymbol{L}'_\theta$, where $\boldsymbol{L}_\theta$ represents a lower triangular matrix having positive values in the diagonal and unrestricted in the other entries. Moreover, some other restrictions have to be considered like setting the first element of $\boldsymbol{L}_\theta$ to get the first element of $\boldsymbol{\Sigma}_\theta$ to be one, and consider gamma priors to the diagonal elements of $\boldsymbol{L}_\theta$ to get only positive values.

The third and last way of parametrisation is to model the ability of each individual in each latent dimension as a linear combination of random coefficients and time, which can allow the ability trajectory to cover most of the variability. Then, the ability in dimension f for subject i at time t has the form $\theta_{ift} = \gamma_{if}^{(0)} + \gamma_{if}^{(1)} t$, where $\gamma_{if}^{(0)} \sim N(\mu_{\gamma_0}, \sigma^2_{\gamma_0})$ and $\gamma_{if}^{(1)} \sim N(\mu_{\gamma_1}, \sigma^2_{\gamma_1})$. There are also some restrictions to consider when modelling to attain identifiability such as setting the mean and variance of one random coefficient to be constant. Hence, $\gamma_{if}^{(0)} \sim N(0, 1)$.

5.1.2 Prior Distributions

The prior distributions chosen for the parameters to be considered in the three approaches are detailed in Table 5.1. It has to be mentioned that these priors were the same for each of the 6 factors. The choice of a truncated $\mathrm{N}(1, 1)$ as a prior for the discrimination parameter is motivated by the fact that first it is restricted to be non-negative and that this distribution is less biased than a truncated $\mathrm{N}(0, 1)$ meaning a less informative prior for the item discrimination allowing the previous knowledge to have a more equal probability in any value between 0 and 2.

5.1.3 Estimation Results

This section presents the results obtained from the modeling strategy applied. However, only the outcomes for the AR(1) and Random Effects approaches are presented since the model concerning the Unstructured Covariance did not attain convergence (this will be explained in subsection 5.1.4). Moreover, the 6 latent factors were run simultaneously for each model, but we did not model the dependence between these factors across time.

All the analysis was performed using a **BUGS** (**B**ayesian inference **U**sing **G**ibbs **S**ampling) code. The openbugs function in the **R2WinBUGS** package (Sturtz et al., 2005) and the **BRugs** package (Thomas et al., 2006) were employed to call **OpenBugs** (Spiegelhalter et al., 2015) from the free software R. See Appendix A for details about the code used, which is based on Curtis (2010). We considered 3 chains with length of 10000 iterations and a burn-in phase of 5000. The outcomes were then pooled in a single chain

Table 5.1: Choice of prior distributions for each f latent dimension

Parameters			AR(1)	Unstructured	Random Effects
Discrimination		α_j	N(1, 1) I[$\alpha_j > 0$]	N(1, 1) I[$\alpha_j > 0$]	N(1, 1) I[$\alpha_j > 0$]
Difficulty		d_j	N(0, 1)	N(0, 1)	N(0, 1)
Latent Ability (θ_i)	$\boldsymbol{\mu}_\theta$	$\mu_{\theta_{i1}}$	0	0	-
		$\mu_{\theta_{i2}}$	N(0, 1)	N(0, 1)	-
		$\mu_{\theta_{i3}}$	N(0, 1)	N(0, 1)	-
	$\boldsymbol{\Sigma}_\theta$	σ	1	-	-
		ρ	U(−1, 1)	-	-
		L_{ii}	-	Gamma(1, 1)	-
		$L_{ij\ [i>j]}$	-	N(0, 1)	-
		$\gamma_i^{(0)}$	-	-	N(0, 1)
	$\gamma_i^{(1)}$	$\mu_{\gamma_i^{(1)}}$	-	-	N(0, 1)
		$\tau_{\gamma_i^{(1)}}$	-	-	Gamma(1, 1)

for each parameter and summarised based on the sample average and credibility intervals to obtain the estimates of the parameters.

AR(1) Covariance Structure

A summary of the ρ_f estimates for each dimension $f = 1, 2, \ldots, 6$ is provided in Table 5.2. It is worthwhile to report this parameter since it can be an indicator of the development in the ability of each specific dimension per time point. Thus, positive values of ρ_f will point out an improvement in the ability across time. A credible interval containing zero will indicate no evidence for development over each time period. The abilities of Pretense, Desire, Think and Location Change had a remarkable evolution across time as its values were above 0.5. The same holds for the abilities related to Deceptive Box, Non Verbal and Verbal False Belief. On the other hand, the Narrative dimension did not seem to have any development ($\rho_5 \simeq 0$) at all. This can be explained first by the response patterns of the children (see Fig 3.1) where it does not show any emphasis on any particular response pattern. It also could be because it is not an appropriate test for children around the age of 3 years old, which is the age of the children under study.

Table 5.2: Summary of ρ estimate

Factor	$\bar{\rho}_f$	$Q_{0.025}$	$Q_{0.975}$
Non Verbal FB	0.44	0.22	0.63
Pretense, Desire, Think	0.65	0.43	0.83
Verbal FB	0.37	0.00	0.74
Deceptive Box	0.47	0.08	0.84
Narrative	0.06	-0.86	0.88
Location Change	0.62	-0.16	0.98

Other parameters of our interest are the item difficulty and discrimination parameters. The credibility interval of these are shown in Fig. 5.1 sorted by the difficulty item estimate. As we can notice, the Narrative task is revealed the easiest item, but it has a wide credibility interval showing a high uncertainty about this task. Moreover, its discrimination is very low which shows the inability of the item to distinguish between high and low abilities of each factor. On the other hand, the questions related to Pretense and Desire are also easy with high values of discrimination. It can be said then that Pretense and Desire are well tasks to be taken to children around the age of study to identify groups with high and low abilities in each dimension of Theory of Mind. Furthermore, the most difficult question was the Standard Location Change and because of this it does not have a good estimate of discrimination.

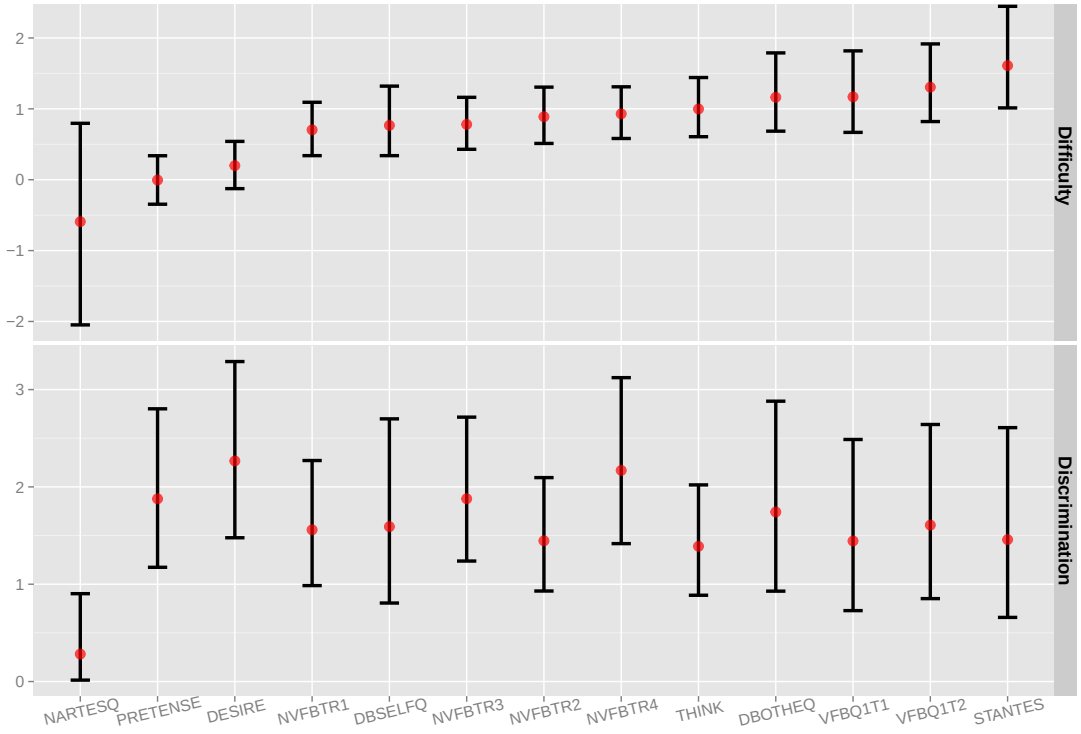

Figure 5.1: Credibility Interval of Item parameters considering AR(1) as covariance structure.

Finally, the latent abilities in each dimension for the 66 children are plotted in Fig.5.2. In general, the ability related to Non Verbal False Belief had an important improvement across time as well as Pretense, Desire and Think. These two have also high scores in the third time point. On the other hand, the latent ability of Deceptive Box goes down for many of the pupils and the Narrative remains constant.

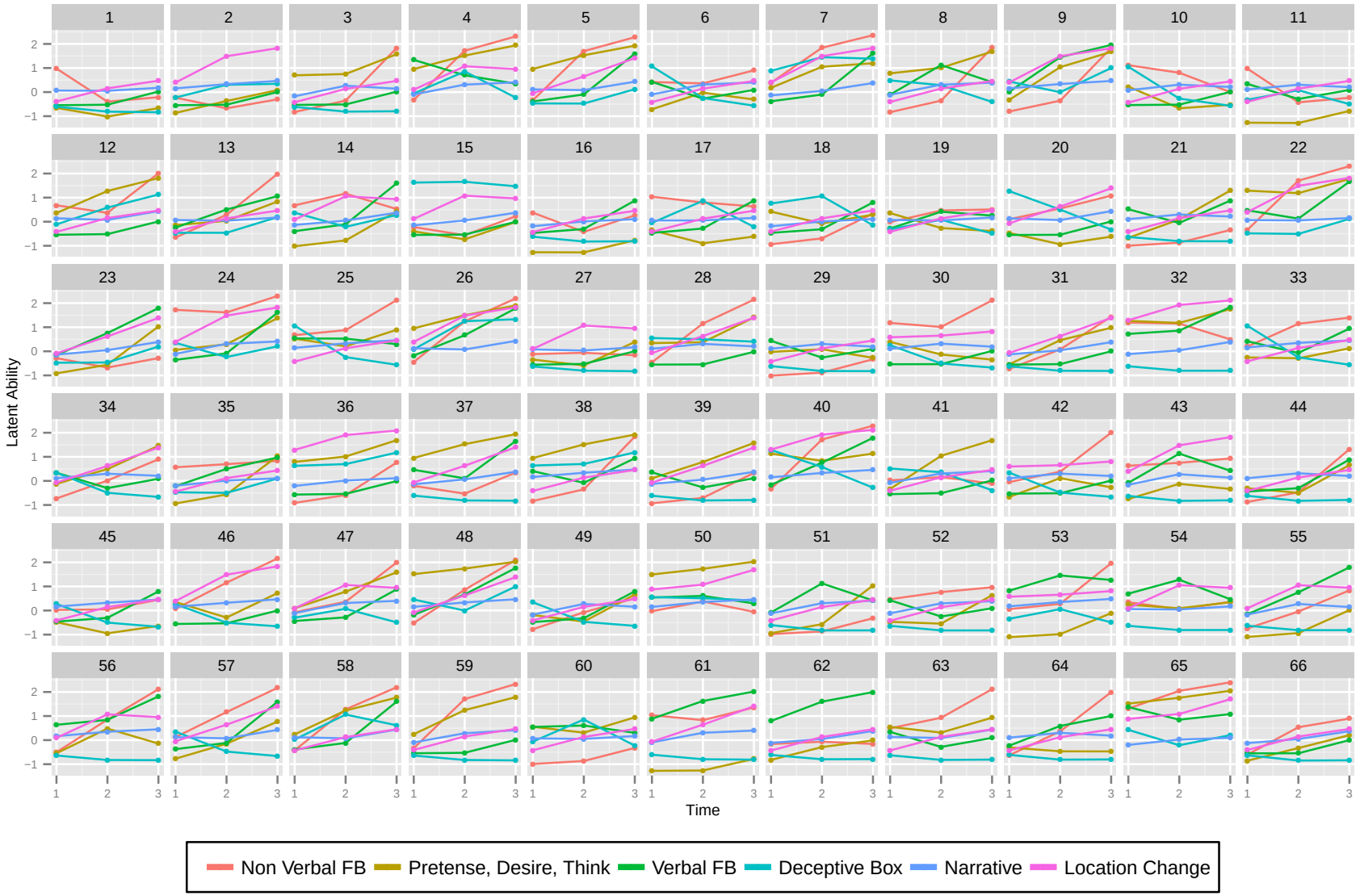

Figure 5.2: Estimated Latent Ability by subject considering the AR(1) as covariance structure.

Random Effects

This model considered the latent ability for each dimension f as a linear combination of time and it contains a random intercept and slope for each individual in every latent factor. These will be estimated besides the item parameters and the latent ability.

Regarding the items parameters of difficulty and discrimination, Fig.5.3 displays the credibility interval of their estimates sorted by item difficulty. It can be noticed that the easiest item was the Narrative proceeded by Pretense and Desire, respectively. However, the estimation is more accurate for the last two, since they have narrower intervals in comparison to the Narrative. Moreover, the Desire item seems to distinguish better between high and low abilities in children since it has the highest discrimination estimate. On the other hand, the second question of Verbal False Belief resulted the most difficult followed by the Standard Location Change item. The Self Deceptive Box item appeared not to be a difficult item (4th place in the list), whereas the Other Deceptive Box item was more difficult than the Non Verbal False Belief trials.

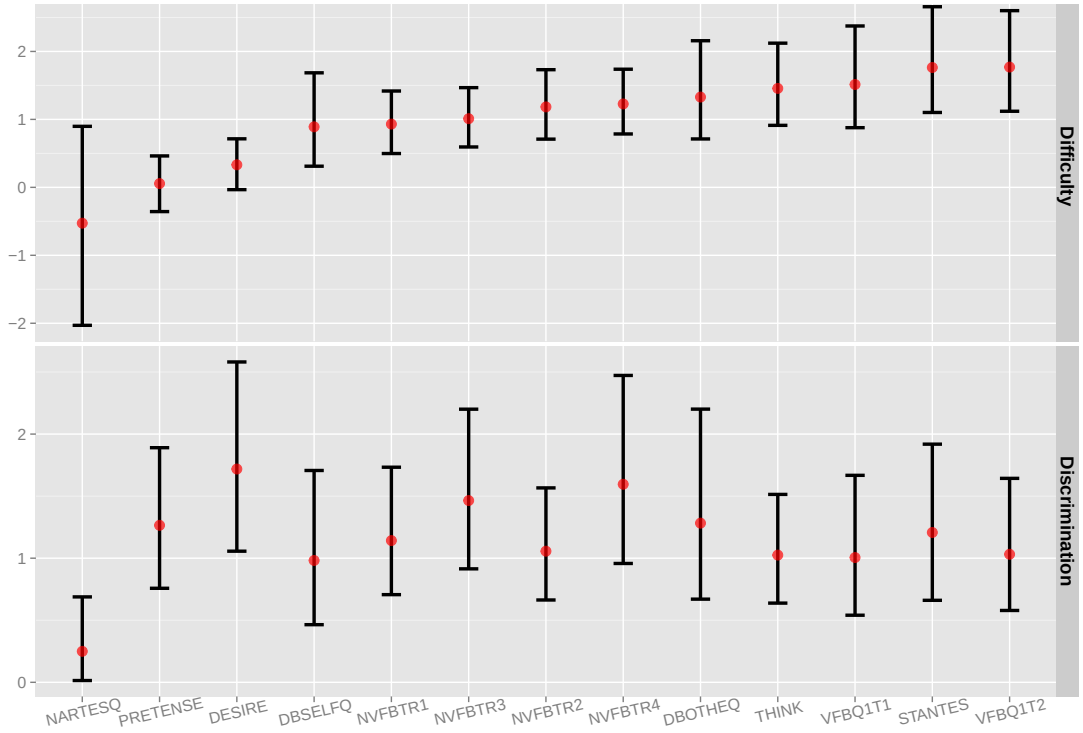

Figure 5.3: Credibility Interval of Item parameters considering random effects. Note the slight difference in ordering in comparison to Fig. 5.1

Another important parameter to show is the slope $\gamma_i^{(1)}$ for each dimension since it can be interpreted as the improvement of the specific ability. Hence, its credibility interval is plotted in Fig. 5.4 sorted by its mean. It can be clearly seen that around 40%

of the children had significant improvement in the ability of Non Verbal False Belief as the credibility interval fall above zero. In relation to Pretense, Desire and Think, about 30% of the children had significant progress in this ability, whereas for Verbal False Belief only around 12% were significant. Even though there were not any significant development in the Deceptive Box ability, it can be seen an increase across time for around 50% of the subjects. On the other hand, regarding the Narrative task, the slopes did no show any trend across time meaning no progress in the ability and also their credibility interval had high uncertainty. For Location Change, the improvement was really small but it was not significant according to the length of the credibility intervals.

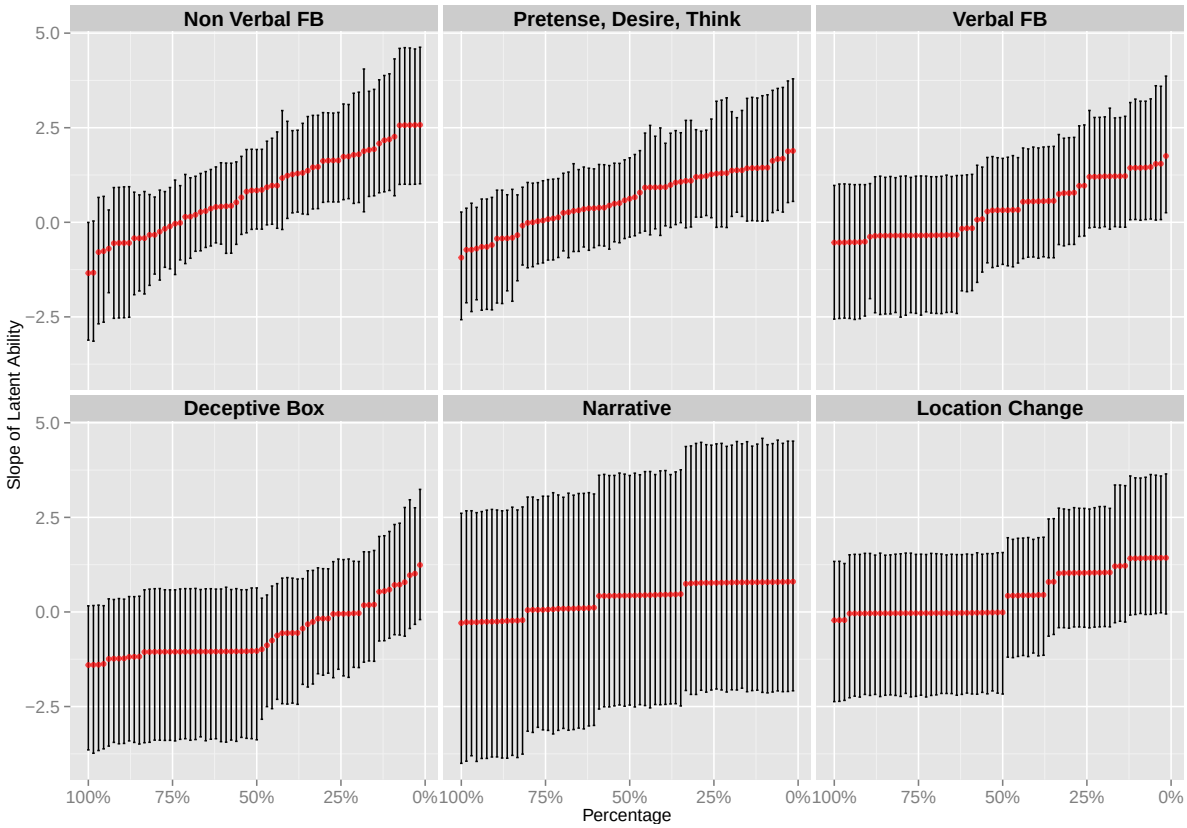

Figure 5.4: Credibility Intervals of the 6 latent abilities slopes.

Lastly, Fig. 5.5 shows the latent abilities in each dimension by individuals. It can be seen that the Non Verbal False Belief ability shows a better improvement across time for most of the children. Similar pattern can be found for the ability related to Pretense, Desire and Think. This can be supported by the fact that the items in those abilities are the easiest according to our previous stated results. On the other hand, the ability of Location Change mostly remains constant or with a very low increase across time. Regarding the Deceptive Box ability, this goes down as times passes for an important number of children. The Verbal False Belief ability had in average a not well learning across time, whereas the Narrative ability had a non specific trend.

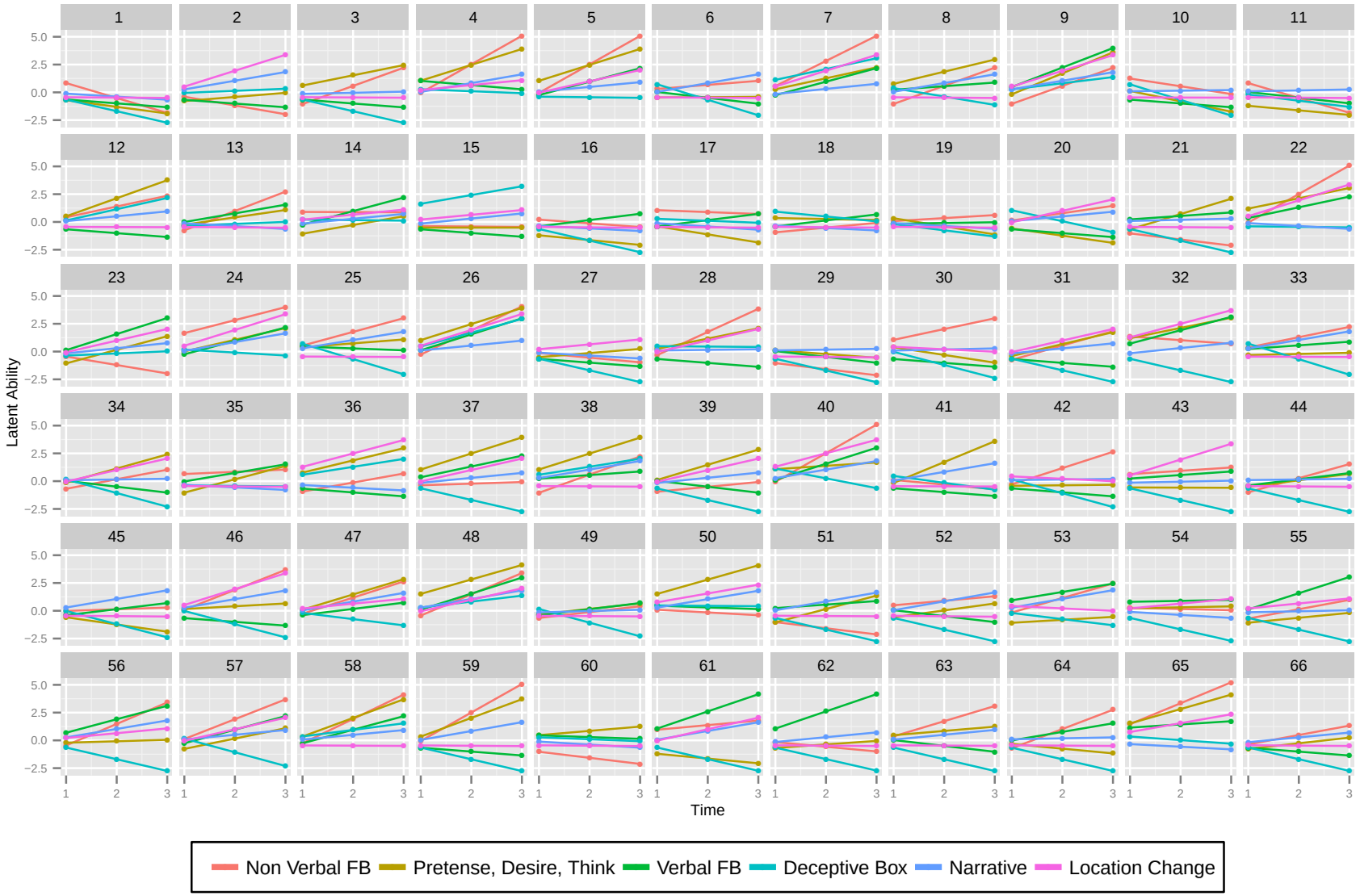

Figure 5.5: Estimated Latent Ability by subject considering random effects.

5.1.4 Convergence Diagnostics

MCMC techniques were used to obtain the estimates of the three models presented in the previous subsection. It is known that when running a MCMC algorithm, it is always needed to analysed the convergence of the chains in order to be sure that the realizations obtained are a sample that come from the stationary distribution. For this purpose, the Gelman Rubin statistics was employed, which is based on the within and between sample variabilities of the chains. The decision rule is that if $\hat{R} < 1.2$ it can be said that convergence is attained.

1. AR(1) Covariance Structure

This model contains the estimation of the item parameters (α, d) and the ability $\boldsymbol{\theta}$, with the latter involving the estimation of ρ and $\boldsymbol{\mu}_\theta$. It can be seen in Fig.5.6 that, according to Gelman Rubin diagnostics, all the parameters had a value less than 1.2 meaning that the chains converged.

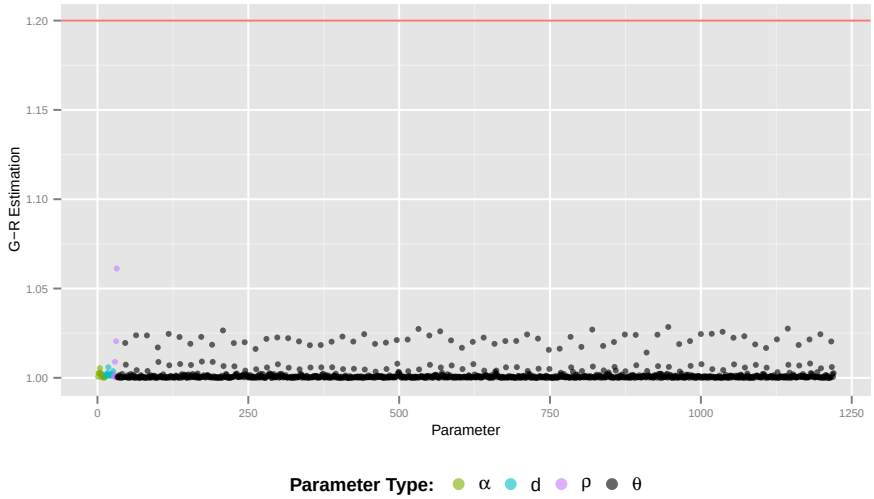

Figure 5.6: Gelman Rubin diagnostic considering AR(1) as covariance structure.

2. Unstructured Covariance

In this model, a non specific structure is assumed for the covariance of the latent ability ($\boldsymbol{\Sigma}_\theta$). Therefore, there are more parameters to be estimated in addition to the general ability $\boldsymbol{\theta}$, the item difficulty (d) and the item discrimination (α). After running the model, the convergence diagnostic was done resulting not significant ($\hat{R} > 1.2$) for a considerable amount of parameters based on the Gelman Rubin statistics, see Fig.5.7. These findings agree with what Curtis (2010) says about this kind of model that in order to get a good sample from the posterior, the chains need to be run considering several number of iterations (e.g. 100000). However, this was not done because of the lack of time. As a result, it does not make sense to continue with the analysis.

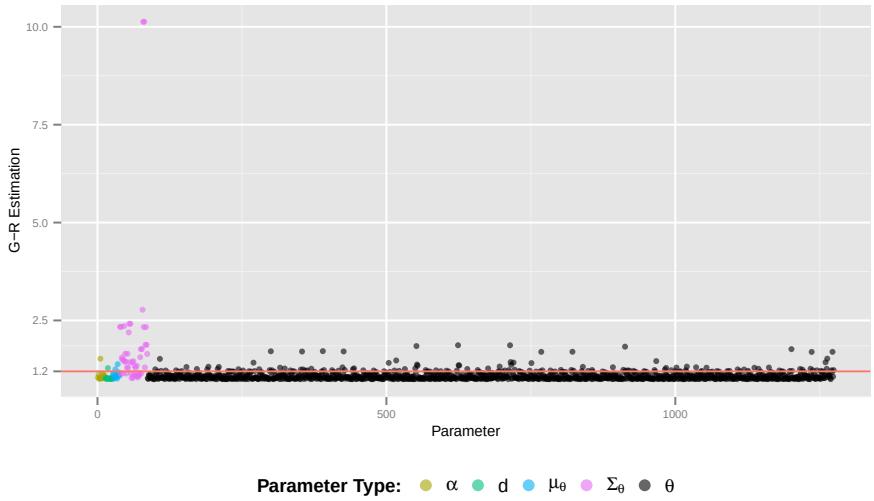

Figure 5.7: Gelman Rubin diagnostic considering unstructured covariance.

3. Random Effects

The last model employed considers the latent ability for each dimension as a linear combination of time and it contains a random intercept and slope for each individual. These will be estimate besides the item parameters and the latent ability. The convergence diagnostics shows that all the parameters converged in the MCMC; that is, the values of the Gelman Rubin statistics resulted less than 1.2 as it is shown in Fig 5.8.

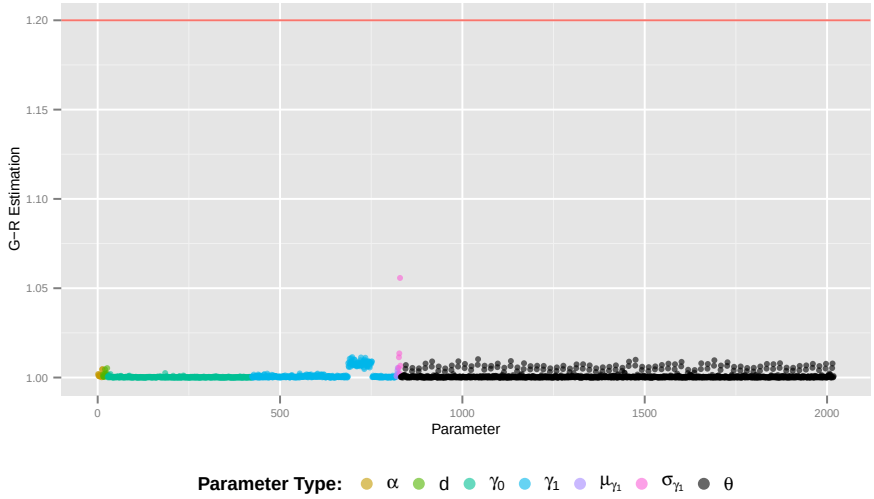

Figure 5.8: Gelman Rubin diagnostic considering random effects.

5.1.5 Model Selection

The three models obtained were compared based on the *deviance information criterion* (DIC) which can be seen as a generalization of the AIC criterion for the bayesian

framework. It is computed as DIC $= \bar{D} + p_D$, where the first component describes the model fitting measured by the posterior expectation of the deviance $\bar{D} = E_{\theta|y}(D(\theta))$, and the second element stands for the complexity of the model measured by the effective number of parameters $p_D = \bar{D} - D(E_{\theta|y}(\theta))$. The decision rule is similar to the AIC; that is, smaller values of DIC suggests a better model.

Therefore, considering this criterion, Table 5.3 shows the values and credibility interval obtained for the three models analysed. Even though, the lowest value went for the Unstructured Covariance approach, it will no be taken into account for the comparison since it did not convergence and its results can be biased as well as its DIC value. Hence, comparing the two remaining approaches, we chose the AR(1) as the best model since it had a lower DIC.

Table 5.3: Summary of DIC criterion

Model	DIC	$\mathbf{Q_{0.025}}$	$\mathbf{Q_{0.975}}$
AR(1) Covariance Structure	2312.46	2205.88	2418.96
Unstructured Covariance	2242.62	2124.69	2359.80
Random Effects	2337.56	2258.15	2415.93

5.2 Second Stage: Ability Regression

In this stage, we take the mean estimates of the 6 dimensional ability score and continue with the final analysis. First, we compare the correlation between those factors at each time point $t = 1, 2, 3$. Second, we regress the ability of each factor at times $t = 2, 3$ against all the other abilities at the previous time step. These two procedures were done considering the AR(1) model because this was chose as the best model according to the deviance information criterion (DIC).

5.2.1 Correlation Analysis

The sample based correlation of the abilities of each dimension across time is displayed on Fig. 5.9. In Time 1, we can see that most of the factors of Theory of Mind are not strongly associated like Non Verbal False Belief with Pretense, Desire and Think. However, the latter has a mild correlation with Location Change at this time point ($\text{corr}(f_2, f_6) = 0.35$). As the time goes by some associations increase for most of the factors being more remarkable for Deceptive Box with Pretense, Desire and Think at time 3 ($\text{corr}(f_1, f_4) = 0.44$). Moreover, an important pattern found is the moderate correlation above 0.32 of Location Change dimension with the factors of Verbal False Belief, Deceptive Box and Pretense, Desire and Think across time.

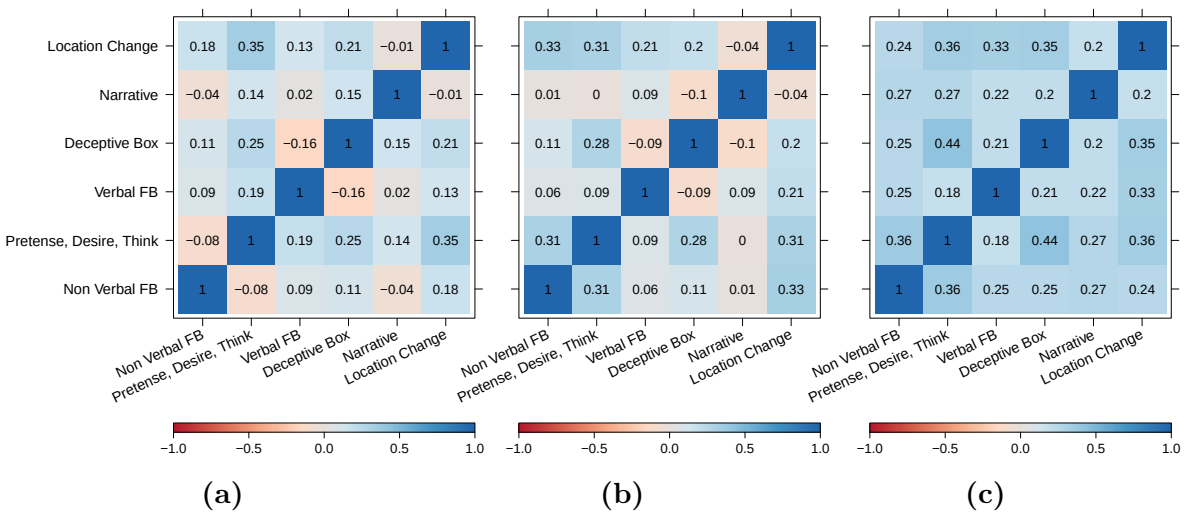

Figure 5.9: Correlations between latent abilities for (a) Time 1, (b) Time 2 and (c) Time 3 resulted from the AR(1) model.

5.2.2 Causal Analysis

We have seen in the previous section that there are some latent abilities that are moderately associated across time. Thus, it results sensible to find how significant the relation is by regressing each dimension ability at times $t = 2, 3$ with the abilities at the preceding times. Hence, each latent ability at time 2 was regressed with all the abilities at time 1 including itself. Similarly was done for time 3 against time 2. Even though, the longitudinal analysis was done in a bayesian framework, for each model perform in this section, we only use a classical linear approach.

Fig. 5.10 displays the results of this phase as a path diagram of the relation across time between the six latent abilities. The red shaded circles stands for the goodness of fit, which in linear models is the R^2, and it varies according to its value. Additionally, the darker the arrow is, the more significant becomes the precedent ability in explaining the respective latent ability at this time point. Considering this, it can be clearly seen that Pretense, Desire and Think can be explained at Time 2 by the same abilities and Non Verbal False Believe abilities had at Time 1. The same pattern is looked for Non Verbal False Belief at Time 2 which is summarised by its previous ability and the Pretense, Desire and Think ability at Time 1. However, these last causality is less accurate since its R^2 value is smaller. Even though there is an arrow from Location Change ability at Time 1 to Verbal False Belief ability at Time 2 indicating causality, it is not that much significant ($p - value \approx 0.05$) and also the R^2 is around 50 %. The remaining abilities at Time 2 are as well explained by themselves at Time 1.

At Time 3, all the abilities are caused by their previous time point value. The remarkable pattern is shown by Pretense, Desire and Think at Time 2 explaining three dimensions at Time 3 apart form itself: Non Verbal False Belief, Dceptive Box and Loca-

tion Change. These last with not a strong p-value but it still having a good precision (R^2 around 0.8). Moreover, Non Verbal False Belief ability seem to be explained also by the Narrative factor, but the significance is almost at the boundary and can be because the p-values, in general, have not been adjusted for multiple comparison.

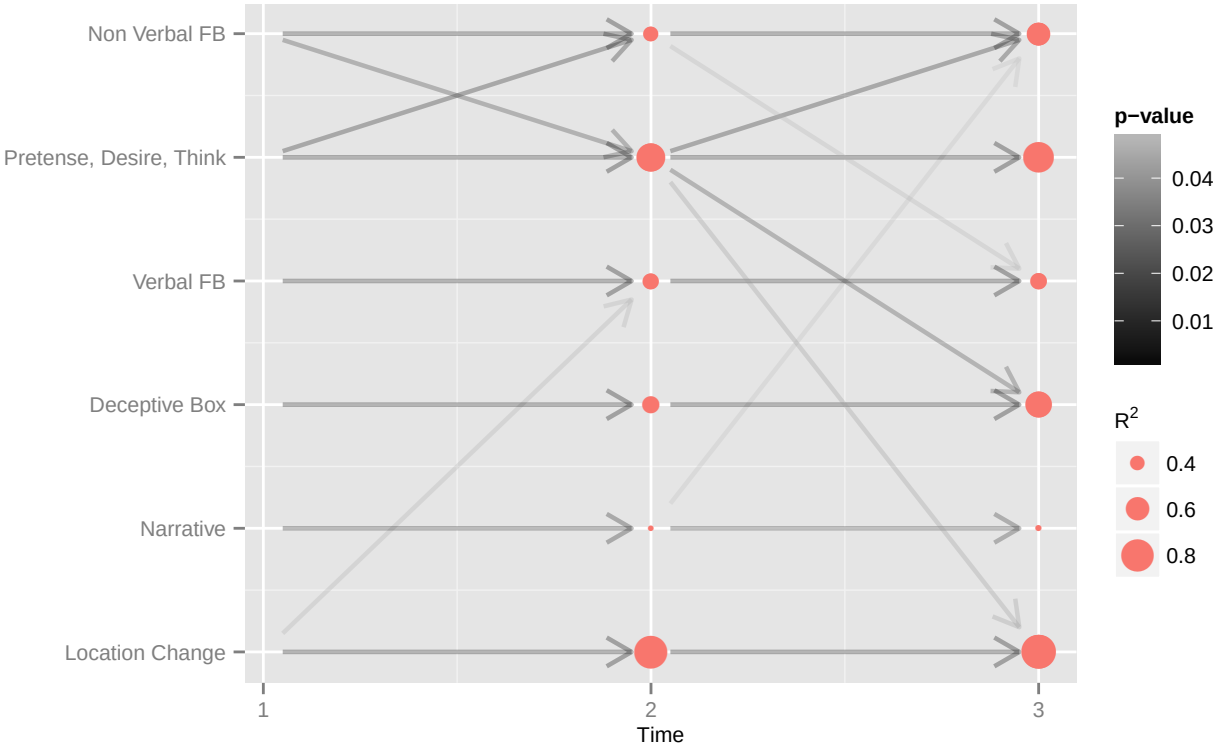

Figure 5.10: Path Diagram of Causality. The p-values have not been adjusted for multiple comparison.

5.3 Conclusion

In this chapter, the main findings were stated from the longitudinal and causal analysis. In the first phase, three approaches were used, but it turned out that the AR(1) structure covariance was the best model according to the DIC criteria. Moreover, it was found out that the easiest item was the Narrative task, but it did not show any improvement across time. The second in the list of easiness were Pretense and Desire items. In addition, the four trials of the Non Verbal False Belief dimension had the best development along time. For the causal analysis, the main finding was the central role of Pretense, Desire and Think that came into play from Time 2 to Time 3.

Chapter 6

Conclusion and Further Work

6.1 Psychological Context

This study showed that children before 4 years old were able to pass some mental states tasks commonly used to assess the acquisition of Theory of Mind ability and that this evolved across the period of study. Specifically, the more relevant tasks for these age were Pretense, Desire and Non Verbal False Belief. These results support the ideas behind the studies of Call and Tomasello (1999) and Lillard and Flavell (1992), in which the former believe that younger children will perform more sensitively if the task does not involve too much linguistic abilities and the latter suggests that Pretense and Desire tasks are able to be passed at ages of 2 years and 5 months. Furthermore, to find any regularities at this age is not predicted by most theoreticians who assume that young three year olds would simply guess on all the tasks except the Desire and Pretense tasks. Even Call and Tomasello (1999) assume that Non Verbal False Belief emerges at the same age as the Verbal version - i.e. at just over 4 years of age. Thus, the progress shown in these measures is interesting to developmental psychologists, as is the use of Item Response Theory approaches.

We successfully reduced the dimension of the general ability Theory of Mind into 6 latent abilities by applying the Bifactor Model as confirmatory item factor analysis. These latent dimensions were Non Verbal False Belief, Pretense, Desire and Think, Verbal False Belief, Deceptive Box, Narrative and Location Change. However, the last two factors which are basically the same item did not belong to any cluster of items formed by the analysis considered, but did show an association with the general ability. For this reason, they were considered as factors in the final analysis.

The analysis considering the Bayesian longitudinal framework drew important results like the easiness of the items concerning Pretense and Desire and the most difficult item which was Standard Location Change. Moreover, after obtaining the continuous latent ability in each dimension for all the subjects, we could observe that Non Verbal False Belief was the ability with more significant improvement across time. However, this was not the case for the abilities of Deceptive Box and Standard Location Change. Moreover, the Narrative task seemed to have a random pattern with children responding correctly or wrong at non specific time. This non-specific pattern can be explained maybe because the task was not conducted properly or it is not a good task to be taken at this age even though according to Lewis et al. (1994), the task can be passed at 3 years 9 months.

On the other hand, regarding the causal analysis, it can be pointed out that there was an important ability that affects the development of most of the others which is Pretense, Desire and Think. It was found that Time 3 measures are best predicted by Desire, Pretense and Think at Time 2, which is interesting as it shows that more complex means of ascertaining mental states emerge from simpler ones. This result is also of interest as Josef Perner, in particular, argues that there is an intellectual revolution at age 4 when he would predict success at all these tasks simultaneously. In conclusion, these results seem to have found the gradual emergence of a grasp of mental states that is consistent in part with Bartsch and Wellman (1989, 1995) but also with Carpendale and Lewis (2004).

6.2 Methodology Issues

The methodology employed concerning two stages to determine causality appeared to be reasonable. Nevertheless, it could have been better if the main analysis would have been done in only one stage, but because of time constraints we were not able to do this using only a Bayesian or Classical approach.

Something that we could have been done for the random effects model considered in the first stage was allowed the slopes for each latent dimension to have a factor specific mean. On the other hand, the second stage of the Two Stage Model was carried out using linear regression. Because this technique assumes covariates are known without error, the second stage should incorporate this uncertainty, but unfortunately this has not been done. We have repeated the analysis in one stage model obtaining similar results but more conservative. Time did not allow us to report those results, but the modelling idea applied can be stated as:

$$\theta_{3,1:F} = \theta_{2,1:F} \times \mathrm{A}_t + \mathrm{C}$$

where A_t is the transition matrix showing how the abilities depends on abilities at previous time step.

Another weakness was that the correlation between latent abilities was not part of the model. However, this should be incorporated in further work as well since it was found that there is a correlation across time. In this case, the dimension of the abilities to model would have been a 18×18 block diagonal matrix, but the off diagonal is still difficult to know how the structure could be. Moreover, if we recall the exploratory analysis, we could see that there was a lot of guessing going on. Thus, a guessing parameter for each item should be considered in further work and compared with the model found in this work by using the deviance information criterion (DIC).

Finally, in this study we did not use covariates in the modelling, but we have the information of age, sex and institution corresponding to each child. This could also be taken into account in the future employing maybe Multilevel Modelling or Dynamic Latent Trait Models as it is explained in Dunson (2003).

References

Adams, R., Wilson, M., and Wang, W. (1997). The multidimensional random coefficients multinomial logit model. *Applied Psychological Measurement*, 21:1–23.

Andersen, E. (1970). Asymptotic Properties of Conditional Maximum Likelihood Estimators. *Journal of the Royal Statistical Society*.

Astington, J. W. (2001). The future of theory-of-mind research: Understanding motivational states, the role of language, and real-world consequences. *Child development*, 72(3):685–687.

Astington, J. W. and Jenkins, J. M. (1995). Theory of mind development and social understanding. *Cognition and Emotion*, 9:151–165.

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

Bartsch, K. and Wellman, H. M. (1989). Young children's attibution of action to beliefs and desires. *Child Development*, 60(4):946–964.

Bartsch, K. and Wellman, H. M. (1995). *Children talk about the mind*.

Bock, R. D. and Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika*, 46(4):443–459.

Bock, R. D., Gibbons, R., and Muraki, E. (1988). Full-Information Item Factor Analysis. *Applied Psychological Measurement*, 12(3):261–280.

Cai, L. (2010a). High-Dimensional Exploratory Item Factor Analysis by a Metropolis-Hastings Robbins-Monro Algorithm. *Psychometrika*, 75(1):33–57.

Cai, L. (2010b). Metropolis-Hastings Robbins-Monro Algorithm for Confirmatory Item Factor Analysis. *Journal of Educational and Behavioral Statistics*, 35(3):307–335.

Call, J. and Tomasello, M. (1999). A nonverbal false belief task: the performance of children and great apes. *Child development*, 70(2):381–95.

Carpendale, J. I. M. and Lewis, C. (2004). Constructing an understanding of mind : The development of children's social understanding within social interaction. *Behavioral and Brain Sciences*, 27:79–151.

Chalmers, R. P. (2012). mirt: A Multidimensional Item Response Theory Package for the R Environment. *Journal of Statistical Software*, 48(6).

Curtis, S. M. (2010). BUGS code for item response theory. *Journal of Statistical Software*, 36:1–34.

Dunn, J., Brown, J., Slomkowski, C., Tesla, C., and Youngblade, L. (1991). Young children's understanding of other people's feelings and beliefs: individual differences and their antecedents. *Child development*, 62:1352–1366.

Dunson, D. B. (2003). Dynamic Latent Trait Models for Multidimensional Longitudinal Data. *Journal of the American Statistical Association*, 98(463):555–563.

Edwards, M. (2010). A Markov Chain Monte Carlo approach to Confirmatory Item Factor Analysis. *Psychometrika*, 75(3):474–497.

Fox, J.-P. (2010). *Bayesian Item Response Modeling: Theory and Applications*. New York: Springer.

Ghosh, M. (1995). Inconsistent Maximum Likelihood for the Rasch Model. *Statistics and Probability Letters*, 23:165–170.

Gibbons, R. D. and Hedeker, D. R. (1992). Full-Information Item Bi-Factor Analyisis. *Psychometrika*, 57(3):423–436.

Jenkins, J. M. and Astington, J. W. (2000). Theory of Mind and Social Behavior: Causal Models Tested in a Longitudinal Study. *Merrill-Palmer Quarterly*, 46(2):203–220.

Johnson, M. S. (2007). Marginal maximum likelihood estimation of item response models in R. *Journal of Statistical Software*, 20(10):1–24.

Lewis, C., Freeman, N., Hagestadt, C., and Douglas, H. (1994). Narrative access and production in preschoolers' false belief reasoning. *Cognitive Development*, 9:397–424.

Lewis, C., Freeman, N. H., Kyriakidou, C., Maridaki-Kassotaki, K., and Berridge, D. M. (1996). Social Influences on False Belief Access: Specific Sibling Influences or General Apprenticeship? *Child Development*, 67(6):2930–2947.

Lillard, A. S. and Flavell, J. H. (1992). Young children's understanding of different mental states.

Lunn, J. A. (2006). *Very young preschoolers' understanding of the mind: a gradual approach?* PhD thesis, Lancaster University.

Perner, J., Leekam, S. R., and Wimmer, H. (1987). Three-year-olds' difficulty with false belief: The case for a conceptual deficit. *British Journal of Developmental Psychology*, 5(2):125–137.

Perner, J., Ruffman, T., and Leekman, S. R. (1994). Theory of mind is contagious: You can catch it from your sibs. *Child Development*, 65(4):1228–1238.

Reckase, M. (2009). *Multidimensional Item Response Theory*. New York: Springer.

Revelle, W. (2015). The psych package Version 1.5.8. URL `https://cran.r-project.org/web/packages/psych/psych.pdf`.

Rodrigues, M., Pelisson, M., Silveira, F., Ribeiro, N., and Silva, R. (2015). Evaluation of theory of mind : A study with students from public and private schools. *Estudos de Psicologia*, 32(2):213–220.

Shakoor, S., Jaffee, S. R., Bowes, L., Ouellet-Morin, I., Andreou, P., Happé, F., Moffitt, T. E., and Arseneault, L. (2012). A prospective longitudinal study of children's theory of mind and adolescent involvement in bullying. *Journal of Child Psychology and Psychiatry*, 53(3):254–261.

Spiegelhalter, D., Thomas, A., Best, N., and Lunn, D. (2015). OpenBUGS Version 3.2.3. URL = http://www.openbugs.net/w/FrontPage.

Sturtz, S., Ligges, U., and Gelman, A. (2005). R2WinBUGS: A Package for Running WinBUGS from R. *Journal of Statistical Software*, 12(3):1–16. URL `http://www.jstatsoft.org/v12/i03/`.

Tavares, H. R. and Andrade, D. F. (2006). Item response theory for longitudinal data: Item and population ability parameters estimation. *Test*, 15(1):97–123.

Thomas, A., O'Hara, B., Ligges, U., and Sturtz, S. (2006). Making BUGS Open. *R News*, 6(1):12–17. URL `http://cran.r-project.org/doc/Rnews/`.

Titman, A. C., Lancaster, G. A., and Colver, A. F. (2013). Item response theory and structural equation modelling for ordinal data: Describing the relationship between KIDSCREEN and Life-H. *Statistical Methods in Medical Research*, pages 1–33.

Wellman, H. M. (1990). *The child's theory of mind*. MIT Press, Cambridge, MA.

Wellman, H. M., Cross, D., and Watson, J. (2001). Meta-Analysis of Theory-of-Mind Development: The Truth about False Belief. *Child Development*, 72(3):655–684.

Wellman, H. M. and Liu, D. (2004). Scaling of Theory-of-Mind Tasks. *Child Development*, 75(2):523–541.

Wimmer, H. and Perner, J. (1983). Beliefs about beliefs: Representation and constraining function of wrong beliefs in young children's understanding of deception. *Cognition*, 13:103–128.

Wirth, R. J. and Edwards, M. C. (2007). Item Factor Analysis: Current Approaches and Future Directions. *Psychological methods*, 12(1):58–79.

Appendices

Appendix A OpenBugs Code

A.1 AR(1) Covariance Structure

```
model {
  # I. Defining the model
  for(q in 1:Z){
    Y[q] ~ dbern(prob[q])
    logit(prob[q]) <- alpha[j[q]]*(theta[i[q],f[q],t[q]]-delta[j[q]])
  }

  # 1.1. Priors on item parameters
  for (jj in 1:n){
    alpha[jj] ~ dnorm(1, 1) I(0, )
    delta[jj] ~ dnorm(0, 1)
  }

  # II. Distribution of latent abilities
  for(ff in 1:F){
    for (ii in 1:N){
      theta[ii,ff, 1:T] ~ dmnorm(mu.theta[ff,], Pr.theta[ff,,])
    }
  }

  # 2.1. Priors for mu.theta of latent abilities
  for(ff in 1:F){
    mu.theta[ff,1] <- 0.0
    for (tt in 2:T){
      mu.theta[ff,tt] ~ dnorm(0, 1)
    }
  }

  # 2.2. Pr.theta (Sigma.theta AR1 structure) of latent abilities
  for(ff in 1:F){
    sigsq.theta[ff] <- 1.0
    Sigma.theta[ff,1, 1] <- 1.0
    for (ii in 2:T){
      Sigma.theta[ff,ii, ii] <- sigsq.theta[ff]
      for (jj in 1:(ii-1)){
        Sigma.theta[ff,ii, jj] <- sigsq.theta[ff]*pow(rho[ff],ii - jj)
        Sigma.theta[ff,jj, ii] <- Sigma.theta[ff,ii, jj]
      }
    }
    Pr.theta[ff,1:T, 1:T] <- inverse(Sigma.theta[ff,,])
  }

  # 2.2.1. Prior for rho
  for(ff in 1:F){
    rho[ff] ~ dunif(-1.0, 1.0)
    }
}
```

A.2 Unstructured Covariance

```
model {
  # I. Defining the model
  for(q in 1:Z){
    Y[q] ~ dbern(prob[q])
    logit(prob[q]) <- alpha[j[q]]*(theta[i[q],f[q],t[q]]-delta[j[q]])
  }

  # 1.1. Priors on item parameters
  for (jj in 1:n){
    alpha[jj] ~ dnorm(1, 1) I(0, )
    delta[jj] ~ dnorm(0, 1)
  }

  # II. Distribution of latent abilities
  for(ff in 1:F){
    for (ii in 1:N){
      theta[ii,ff, 1:T] ~ dmnorm(mu.theta[ff,], Pr.theta[ff,,])
    }
  }

  # 2.1. Priors for mu.theta of latent abilities
  for(ff in 1:F){
    mu.theta[ff,1] <- 0.0
    for (tt in 2:T){
      mu.theta[ff,tt] ~ dnorm(0, 1)
    }
  }

  # 2.2. Prior for Pr.theta (Sigma.theta unstructured) of latent abilities
  for(ff in 1:F){
    L.theta[ff,1, 1] <- 1.0
    for (ii in 2:T){
      L.theta[ff,ii, ii] ~ dgamma(1, 1)
      for (jj in 1:(ii-1)){
        L.theta[ff,ii, jj] ~ dnorm(0, 1)
        L.theta[ff,jj, ii] <- 0.0
      }
    }
    for (ii in 1:T){
      for (jj in 1:T){
        Sigma.theta[ff,ii, jj] <- inprod(L.theta[ff,ii, 1:T], L.theta[ff,jj, 1:T])
      }
    }
    Pr.theta[ff,1:T,1:T]<-inverse(Sigma.theta[ff,,])
  }
}
```

A.3 Random Effects

```
model {
  # I. Defining the model
  for(q in 1:Z){
    Y[q] ~ dbern(prob[q])
    logit(prob[q]) <- alpha[j[q]]*(theta[i[q],f[q],t[q]]-delta[j[q]])
  }

  # 1.1. Priors on item parameters
  for (jj in 1:n){
    alpha[jj] ~ dnorm(1, 1) I(0, )
    delta[jj] ~ dnorm(0, 1)
  }

  # II. Latent ability model: Intercepts and slopes
  for(ff in 1:F){
    for (ii in 1:N){
      for (tt in 1:T){
        theta[ii,ff,tt] <- gamma0[ff,ii] + gamma1[ff,ii]*(tt-1)
      }
    }
  }

  # 2.1. Priors on intercepts and slopes
  for(ff in 1:F){
    for (ii in 1:N){
      gamma0[ff,ii] ~ dnorm(0.0, 1.0)
      gamma1[ff,ii] ~ dnorm(mu.gamma1[ff], pr.gamma1[ff])
    }
    mu.gamma1[ff] ~ dnorm(0.0, 1.0)
    pr.gamma1[ff] ~ dgamma(1.0, 1.0)
    sigsq.gamma1[ff] <- 1.0/pr.gamma1[ff]
  }
}
```

Appendix B R Procedure

This appendix reports the procedure used in R software to run the **OpenBugs** code.

```
# Run models with Bugs ------------------------------------------------
library(R2WinBUGS)
library(BRugs)

# 1. Prepare data -----------------------------------------------------
# main variables
Y = dataBugs$Y # answer
i = dataBugs$i # subject
j = dataBugs$j # item
t = dataBugs$t # time
f = dataBugs$f # factor
# constants
Z = nrow(dataBugs)   # number of observations
N = length(unique(i)) # number of subjects
n = length(unique(j)) # number of items
T = length(unique(t)) # number of times
F = length(unique(f)) # number of factors
# data
data <- list('Y', 'i', 'j', 't', 'f', 'Z','N', 'n', 'T', 'F')
```

```
20
21 # 2. AR(1) covariance structure: Constant variance ----------------------
22 # set options
23 monitor = c('theta', 'alpha','delta','rho','mu.theta')
24 bugs.file = '1_six_AR.bug'
25 inits = function(){
26   list(alpha = abs(rnorm(n, 0, 1)),
27        delta = rnorm(n, 0, 1),
28        mu.theta = matrix(c(rep(NA, F), rnorm((T - 1)*F)), ncol = T),
29        rho = runif(F))
30 }
31
32 # run model
33 system.time(
34   openbugsoutf6_AR <- openbugs(data = data, inits = inits, model.file = bugs.file,
35                                parameters = monitor, n.chains=3, n.iter=10000,
36                                n.thin=1, n.burn=5000))
37
38 # 3. Unstructured covariance ------------------------------------------------
39 # set options
40 monitor = c('theta', 'alpha', 'delta', 'mu.theta', 'Sigma.theta')
41 bugs.file = '2_six_unstruct.bug'
42 inits = function(){
43   L.theta = array(NA, c(F, T, T))
44   for (ff in 1:F) {
45     L.theta[ff,,][lower.tri(L.theta[ff,,])] = rnorm(choose(T, 2))
46     diag(L.theta[ff, , ]) = rgamma(T, 1.0, 1.0)
47     L.theta[ff, 1, 1] = NA
48   }
49   list(alpha = abs(rnorm(n, 0, 1)),
50        delta = rnorm(n, 0, 1),
51        mu.theta = matrix(c(rep(NA,F), rnorm((T - 1)*F)), ncol = T),
52        L.theta = L.theta)
53 }
54
55 # run model
56 system.time(
57   openbugsoutf6_UNS <- openbugs(data = data, inits = inits, model.file = bugs.file,
58                                 parameters = monitor, n.chains = 3, n.iter = 10000,
59                                 n.thin = 1, n.burn = 5000))
60
61 # 4. Slope and intercept random effects -------------------------------------
62 # set options
63 monitor <- c('theta', 'alpha', 'delta', 'gamma0', 'gamma1', 'mu.gamma1', 'sigsq.gamma1')
64 bugs.file <- '3_six_ME.bug'
65 inits <- function(){
66   list(alpha = abs(rnorm(n, 0, 1)),
67        delta = rnorm(n, 0, 1),
68        gamma0 = matrix(rnorm(N*F, 0, 1), ncol = N),
69        gamma1 = matrix(rnorm(N*F, 0, 1), ncol = N)
70   )
71 }
72
73 # run model
74 system.time(
75   openbugsoutf6_ME <- openbugs(data = data, inits = inits, model.file = bugs.file,
76                                parameters = monitor, n.chains = 3, n.iter = 10000,
77                                n.thin = 1, n.burn = 5000))
```