Statistical models of children’s strategy change in analogical reasoning

Sanne Cnossen
Abstract

This study addresses two research objectives. The first objective was to investigate whether training and age related changes in strategy use were present with regard to solving figural analogical problems. Six analogical reasoning experiments were conducted with a total of 1007 school children participating ($M = 7.3$ years, $90\%$ range $5.2 – 10.2$). Each experiment had a pretest-training-posttest design. The children were randomly allocated to one out of four training conditions: graduated prompts ($N = 431$), outcome feedback ($N = 202$), practice ($N = 279$) or control ($N = 95$). The second objective was to find the most appropriate polytomous IRT model suitable for the analyses of the current dataset. Three models were investigated: the partial credit model (PCM), graded response model (GRM) and the continuation ratio model (CRM). Based upon fit indices, interpretation of the parameters and substantial features of the data, the GRM was selected as the most appropriate model. This model was then used to investigate the sources of individual differences in initial ability and performance change in strategy use from pretest to posttest.

Results showed that age was a significant predictor of analogical reasoning skills. Older children were found to have higher initial ability scores than younger children. In addition, younger children showed greater improvement from pretest to posttest. Graduated prompts trained children showed significantly more improvement compared to children trained in the control, practice and outcome feedback condition. Interaction effects between training condition and age showed younger children to benefit more than older children from the graduated prompts training compared to the other training conditions.
Acknowledgements

This master thesis is the final project of the master program ‘Methodology and Statistics in Psychology’. After finishing the master program ‘Child and Adolescent Psychology’ at the Universiteit Leiden, my enthusiasm and interest in research and statistics had grown. I am thankful for the opportunity that I could extent my study with a second master’s degree. This master was very interesting and valuable for my employment opportunities in the near future.

While conducting the research for my master thesis, I was fortunate to have the guidance of two very good supervisors. First of all, I would like to thank Dr. Claire Stevenson for providing me with interesting research ideas and data for the realization of this thesis. Also, your thorough and clear comments on the research proposal and previous versions were very helpful. To combine both substantial and methodological questions and considerations in one thesis was exciting. Hopefully this thesis will provide you with intriguing thoughts for further research.

Second, I would like to thank Prof. Dr. Henk Kelderman for providing the supervision during Claire’s maternity leave. Your support helped me to gain confidence in my knowledge and ability of statistics and methodology as well as my personal development. Our meetings and conversations were inspiring and enjoyable and your advice and guidance was invaluable. It was a rewarding and excellent closure of my academic carrier at Leiden University.
## Content list

1. **Introduction** .................................................................................................................................................. 5  
   1.1 Analogical reasoning.......................................................................................................................................... 5  
   1.2 Development of analogical reasoning in children .......................................................................................... 5  
   1.3 Solution strategies in analogical reasoning problems .................................................................................... 6  
   1.4 Appropriate methods for the study of ordinal polytomous answer categories ......................................... 7  
       1.4.1 Adjacent Category Models ...................................................................................................................... 8  
       1.4.2 Cumulative Probability Models ............................................................................................................... 10  
       1.4.3 Continuation Ratio Models ................................................................................................................... 12  
   1.5 Appropriate method for the study of repeated measures data .................................................................... 15  
2. **Research questions** ......................................................................................................................................... 17  
3. **Method**............................................................................................................................................................ 18  
   3.1 Design and procedure ......................................................................................................................................... 18  
   3.2 Participants ....................................................................................................................................................... 18  
   3.3 Material ............................................................................................................................................................ 19  
       3.3.1 Figural analogy task ................................................................................................................................. 19  
       3.3.2 Pre- and posttest ........................................................................................................................................ 20  
       3.3.3 Training .................................................................................................................................................... 22  
   3.4 Variables .......................................................................................................................................................... 22  
       3.4.1 Response variable .................................................................................................................................... 22  
       3.4.2 Person predictors .................................................................................................................................... 23  
   3.5 Structural model ............................................................................................................................................... 24  
   3.6 Trait level distributions per experiment ....................................................................................................... 25  
   3.7 Model selection ................................................................................................................................................ 25  
   3.8 Statistical analyses ......................................................................................................................................... 26  
       3.8.1 Software .................................................................................................................................................. 26  
4. **Results** ............................................................................................................................................................. 28  
   4.1 Psychometric properties ................................................................................................................................. 28  
       4.2 Proportion of strategy use .......................................................................................................................... 28  
   4.3 Methodological research question .................................................................................................................. 29  
       4.3.1 Partial Credit Model ............................................................................................................................... 30  
       4.3.2 Graded Response Model ....................................................................................................................... 31  
       4.3.3 Continuation Ratio Model ..................................................................................................................... 32
4.3.4 Multiple groups .................................................................................................................. 34
4.3.5 Model selection .................................................................................................................. 35
4.3.6 Conclusion methodological research question ................................................................. 38
4.4 Substantial research question ............................................................................................. 39
  4.4.1 Proportion of strategy use per training condition ........................................................... 39
  4.4.2 Comparability of the training conditions ....................................................................... 40
  4.4.3 Embretson’s Approach .................................................................................................... 41
  4.4.4 Latent regression ............................................................................................................. 42
  4.4.5 Null model ...................................................................................................................... 42
  4.4.6 Person predictors ............................................................................................................ 43
  4.4.7 Final model and interpretation ....................................................................................... 45
5. Discussion ............................................................................................................................... 47
  5.1 Model selection ................................................................................................................... 47
  5.2 Effect of training and age on the change of strategy use .................................................... 48
  5.3 Limitations ......................................................................................................................... 49
  5.4 Methodological considerations .......................................................................................... 49
  5.5 Recommendations for future research .............................................................................. 50
  5.6 Conclusions ........................................................................................................................ 51
6. Literature ............................................................................................................................... 53
Appendix ........................................................................................................................................ 57
1. Introduction

1.1 Analogical reasoning

Analogical reasoning involves solving problems by identifying corresponding structures in the comparison of known objects and events, and using those structures to gain understanding of a new concept (Goswami, 1992; Siegler & Alibali, 2005). In general, analogical reasoning belongs to the category of inductive reasoning, which has often been related to general intelligence (Csapó, 1997). Inductive reasoning, and especially analogical reasoning, is a way of transferring meaning that knowledge obtained in one context will be applied in a new situation. It is regarded as an important component in the development of children’s cognition (Richland, Morrison & Holyoak, 2006) and an essential skill for school learning (Goswami, 1992). For these reasons, analogical reasoning has been the focus of much research over the years.

1.2 Development of analogical reasoning in children

The development of analogical reasoning is considered to be a very important concept for investigation since it provides insight into children’s intellectual capacities (Stevenson, Touw & Resing, 2011). For many years, researchers differed in their opinion whether the ability to reason by analogy was present in young children. Nowadays, most researchers assume that this ability is indeed present since research showed analogical reasoning abilities in progressively younger children (Goswami, 1991; Singer-Freeman, 2005). Additional to the presence of analogical reasoning skills in young children, Goswami (1991) stated that later on in childhood, also qualitative developments occur. Age is thus an important factor in analogical reasoning abilities. Research has shown that older children are better at solving analogical problems than younger children (e.g., Hosenfeld, van den Boom & Resing, 1997).

A very important concept in the development of analogical reasoning is the strategic development (Siegler, 1999). The majority of the studies that investigate the ability to reason by analogy use the correctness of the responses. The strategies used to derive these responses are less often investigated while they provide interesting information about whether and how a child is able to reason by analogy. The strategic development of a child gives insight into a child’s learning (Siegler, 1999) and is therefore relevant to educational psychologists and teachers (Stevenson et al., 2011). As described in Tunteler and Resing (2007b), based on other research, children do not simply replace a less appropriate strategy with a more appropriate one.
The changes and shifts in cognitive strategy use appear to be a complex process that occurs gradually (Siegler, 1999).

### 1.3 Solution strategies in analogical reasoning problems

Tunteler and Resing (2007b) studied the effects of practice on the development of spontaneous analogical transfer in 5-8 year old children from story problems to physical tasks. They distinguished three groups of children with different reasoning strategies; 1) children who consistently show analogical reasoning over trials; 2) children who consistently show inadequate, non-analogical reasoning; and 3) children who show varying reasoning (adequate as well as inadequate). The results showed that, regardless of age, the ability to use analogical transfers spontaneously improved with practice. Large individual differences were found regarding the problem-solving strategies. Children used adequate as well as inadequate strategies at the same time, which might be evidence for a continuous and gradual, quantitative change process in the development of analogical problem solving (Tunteler & Resing, 2007b).

Tunteler, Pronk and Resing (2007a) studied the changes of children’s abilities on geometric analogical reasoning problems and the additional effect of a short training procedure. A total of 36, 6-8 year-old first-grade children participated in the study. They distinguished four types of strategies, namely 1) explicit correct analogical; 2) implicit correct analogical; 3) incomplete analogical; and 4) non-analogical associative. The difference between explicit and implicit strategies was that with an explicit strategy, the child had explicitly named all the transformations that the item contained. This in contrast to an implicit strategy whereby all the transformations were present, but not explicitly stated by the child. Results showed that repeated practice led to an improvement of, mainly implicit, analogical reasoning. Training led to even more improvement, mostly due to an increase in explicit analogical reasoning. In line with other research, there was a relative large number of children who showed a gradual change in using analogical reasoning strategies and a relative smaller number of children showed a more rapid change.

Stevenson et al. (2011) investigated whether the learning and strategy progression of analogical reasoning skills of children followed similar patterns regardless of the assessment mode (paper-based and computerized). They classified strategies as 1) correct analogical; 2) partial analogical with one or two incorrectly applied transformations; 3) duplication; and 4) other non-analogical. The progression of children’s solution strategies was measured during weekly sessions over four consecutive weeks. Results showed that, in both assessment
conditions, much variability was found regarding the solution strategies. In both conditions, children used on average more than three strategies within each test session. In addition, a practice effect was found leading to improvement of solution strategy use especially from the first to the second session.

1.4 Appropriate methods for the study of ordinal polytomous answer categories

In line with the studies discussed in the previous paragraph, this study also focuses on the strategies used to answer analogical reasoning problems. Strategy use is a polytomous response variable and therefore requires a model appropriate for the analyses of polytomously scored items. Additionally, strategy use is regarded as an ordinal variable. Item response theory (IRT) models which allow for multiple ordered-response categories per item appear to be appropriate in this situation.

IRT models are used to estimate a person’s trait level based on the person’s responses and the properties of the items that were administered (Embretson & Reise, 2000). With a polytomous IRT model, the nonlinear relation between the continuous latent trait level and the probability of responding in a particular category is represented. Polytomous responses are handled by forming logits (Rijmen, Tuerlinckx, de Boeck & Kuppens, 2003). A logit is defined as the logarithm of the ratio of the probability of responding in a subset $A$ of all categories, relative to the probability of responding in a disjoint subset $B$ of all categories (De Boeck & Wilson, 2004). There are different ways how the categories are classified into the subsets $A$ and $B$, leading to different logits. Three possible logits for polytomous data are adjacent-categories logits, cumulative logits and continuation-ratio logits (Rijmen et al., 2003). These logits are all appropriate for ordinal responses since ordering information is taken into account (Rijmen et al., 2003). The models this study focuses on are derived from these three different types of logits. They are respectively the partial credit model, graded response model and the continuation ratio model. These models all assume local independence of the item responses and a unidimensional trait level, which are two assumptions required for estimating item parameters with IRT models (Masters, 1982; Samejima, 1969; Hemker, van der Ark & Sijtsma, 2001). The first assumption, local independence, means that the response to an item is unrelated to any other item when controlled for trait level so that trait level explains all relations between item responses. The second assumption is appropriate dimensionality, which means unidimensionality in context of the three models that will be discussed. Unidimensionality
means there is a single latent trait variable sufficient to explain the common variance among item responses (Embretson & Reise, 2000).

For clarification, some indices are explained first. Assume that a test has $I$ items ($i = 1, 2, \ldots, I$) with item $i$ having $h_i = m_i + 1$ response categories. These response categories are indexed as $x$ ($x = 0, 1, \ldots, m_i$) with all values of $x$ being successive integers. The random variable for the chosen category of subject $s$ ($s = 1, 2, \ldots, S$) on item $i$ is denoted by $X_{is}$.

In this study we will examine which of the three IRT models for polytomous data will be best suited for the measurement of children’s analogical reasoning strategies.

1.4.1 Adjacent Category Models

The first class of polytomous IRT models is the class of adjacent category models (ACMs). A well-known model from the ACMs is the partial credit model (PCM) developed by Masters (1982) for the analysis of partial credit data. His model extends the binary Rasch model to the polytomous case. As described in Embretson & Reise (2000), the binary Rasch model is the simplest IRT model. In this model, the dependent variable is a dichotomous response (i.e., 1 or 0 for correct vs. incorrect) of a person to an item. Under the Rasch model, the probability of a correct response of subject $s$ on item $i$ can be expressed as follows:

$$P(X_{is} = 1 | \theta_s, \beta_i) = \frac{\exp(\theta_s - \beta_i)}{1 + \exp(\theta_s - \beta_i)}$$  \hspace{1cm} (1)

where $\theta_s$ denotes the subject’s latent trait level and where in this study, it is assumed that the subject was randomly selected from the population ($\theta \sim N(0, \sigma_\theta^2)$) and $\beta_i$ denotes the item difficulty parameter.

As mentioned, the partial credit model (PCM) is an extension of the Rasch model. The PCM can handle several ordered levels of performance on each item and awards partial credit for partial success on items (Masters, 1982). To illustrate the PCM, an example from Masters (1982) is presented. Suppose an item has four response categories (‘strongly disagree’, ‘disagree’, ‘agree’, ‘strongly agree’). A person who chooses ‘agree’ can be considered to have chosen ‘disagree’ over ‘strongly disagree’ (first step taken) and ‘agree’ over ‘disagree’ (second step taken) but failed to choose ‘strongly agree’ over ‘agree’ (third step rejected) (Masters, 1982). Consequently, multiple steps can be completed in the PCM. This in contrast with the Rasch model where only a single step can be taken, namely from an incorrect answer to a correct answer.
Let the response categories be labelled 0, 1, 2 and 3 with 0 being ‘strongly disagree’ and 3 being ‘strongly agree’. The probability of a person to take the third step in item \( i \) in order to score 3 rather than 2 (if they already reached the second step) is written as follows:

\[
P_{i3}(\theta) = \frac{\exp(\theta - \delta_{i3})}{1 + \exp(\theta - \delta_{i3})} \quad (2.1)
\]

where \( \delta_{i3} \) is defined as the difficulty of the third step in item \( i \).

Similarly to Equation 2.1, the probability of a person to take the second step in item \( i \) in order to score 2 rather than 1 (if they already reached the first step) can be calculated as follows:

\[
P_{i2}(\theta) = \frac{\exp(\theta - \delta_{i2})}{1 + \exp(\theta - \delta_{i2})} \quad (2.2)
\]

The probability of taking the first step in item \( i \) in order to score 1 rather than 0 is identical to the Rasch model presented in Equation 1 except that the difficulty parameter of the item \( (\beta_i) \) is replaced by the difficulty parameter \( \delta_{i1} \) of the first step. So, it is clear that the PCM relies on the adjacent ratio (1 vs. 0, 2 vs. 1 and 3 vs. 2) (see Figure 1) and therefore is an adjacent category model (ACM; Hemker et al., 2001; De Boeck & Partchev, 2012; De Boeck & Wilson, 2004).

Combining the equations into one general expression for the probability of a person scoring \( x \) on item \( i \) results in the PCM:

\[
P_{ix}(\theta) = \frac{\exp\left[\sum_{j=0}^{x}(\theta - \delta_{ij})\right]}{\sum_{y=0}^{\sum_{j=0}^{m_i}}[\exp(\sum_{j=0}^{y}(\theta - \delta_{ij})]\] \quad (3)
\]

with \( \sum_{j=0}^{0}(\theta - \delta_{ij}) \) defined as 0 and where \( \delta_{ij} \) are the category intersections and \( j \) denotes the item steps that can be completed \( (j = 1, 2, \ldots, m_i) \) (Masters, 1982). In this model, the probability is calculated from the exponent of the sum of all \( (\theta - \delta_{ij}) \) terms for each category up to \( x \), divided by the sum of the numerator terms for all possible categories.
The item parameter $\delta_{ij}$ indicates the difficulty of the $j$’th step in item $i$ and is the point on the trait continuum where the probability curves for categories $j-1$ and $j$ intersect. These parameters show where one category becomes more likely than the previous category. This can be seen when the PCM is displayed graphically, by plotting the probabilities of responding in each category as a function of $\theta$, called the category response curves (CRCs). An example of the category response curves of an item with four categories is presented in Figure 2.

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{Figure2.png}
\caption{Category Response Curves of a polytomous scored example item.}
\end{figure}

Since the PCM is a member of the Rasch family it shares a distinguishing characteristic, namely the separability of the parameters (Masters, 1988). This results in a sufficient statistic for the person’s ability, which is the count of the total number of steps the person completed (the raw scale score). For the item parameters the sufficient statistic is the count of the number of persons that have completed each step (Masters, 1982).

### 1.4.2 Cumulative Probability Models

The second class of models is the class of cumulative probability models (CPMs). A well-known model from this class is the graded response model (GRM) developed by Samejima (1969). The GRM is an extension of the two-parameter logistic model (2PL), which includes two parameters to represent item properties (Embretson & Reise, 2000). In addition to the item’s difficulty parameter ($\beta_i$), also the item discrimination parameter ($\alpha_i$) is included in the 2PL model. The item discrimination parameter represents how steeply the rate of success varies
with trait level. In the 2PL model, the probability of a person with trait level $\theta$ to pass a dichotomously scored item $i$ is given as follows:

$$P(X_i = 1|\theta, \beta_i, \alpha_i) = \frac{\exp[\alpha_i(\theta - \beta_i)]}{1 + \exp[\alpha_i(\theta - \beta_i)]}$$

(4)

with, in this study, a normal distribution for the $\theta$'s and means equal to zero ($\theta \sim N(0, \sigma^2_\theta)$).

As mentioned, the graded response model extends the 2PL dichotomous model to the polytomous case. Like the PCM, the GRM is appropriate when item responses are ordered categorical responses. In the GRM, each item is described by the item slope parameter ($\alpha_i$) and $j$ ($j = 1, 2, \ldots, m_i$) between category threshold parameters ($\beta_{ij}$) (Samejima, 1969). Consider the example presented in the previous paragraph with four response categories ($x = 0, 1, 2, 3$). In this example, there are three between category thresholds namely $j = 1, 2, 3$. To derive the conditional probability of responding in a particular response category, two steps have to be taken. The first step concerns the probability of a person’s item response ($x$) to fall in or above a given category threshold ($j$) conditional on trait level ($\theta$). This is given by the following equation:

$$P_{ix}^*(\theta) = \frac{\exp[\alpha_i(\theta - \beta_{ij})]}{1 + \exp[\alpha_i(\theta - \beta_{ij})]}$$

(5)

with $P_{i0}^*(\theta) = 1$ and $P_{im}^*(\theta) = 0$ and where $x = j$.

The item parameters $\beta_{ij}$ in the GRM have a different meaning than the item parameters in the PCM. In the GRM they represent the trait level necessary to respond above threshold $j$ with a .50 probability. Notice that the $\alpha_i$ parameters in the GRM are not referred to as the discrimination parameters. Instead, they are called slope parameters. This is due to the fact that the discrimination of the item also depends on the spread of the category thresholds $j$. In the GRM, an item is treated as a series of $m_i$ dichotomies. In the present example, this means that with a 2PL model, the probabilities ($P_{ix}^*(\theta)$) of $x = 0$ vs. 1, 2 and 3, $x = 0$ vs. 1, 2, 3 and $x = 0$, 1 and 2 vs. 3 (see Figure 3) are calculated with the constraint that the slopes are equal within an item. This shows that GRM is a cumulative probability model (CPM; Hemker et al., 2001).
In the second step, the probability of a person responding in category \( x \) to item \( i \) is obtained by subtracting the cumulative probabilities (Samejima, 1969). Using our four-category example, the probabilities to respond in a certain category are given by equations 6.1 to 6.4. In addition, these equations can be written as one general equation (7) with \( \sum_{x=0}^{m_i} P_{ix}(\theta) = 1 \).

\[
P_{i0}(\theta) = 1 - P_{i1}(\theta) \tag{6.1}
\]

\[
P_{i1}(\theta) = P_{i1}^*(\theta) - P_{i2}^*(\theta) \tag{6.2}
\]

\[
P_{i2}(\theta) = P_{i2}^*(\theta) - P_{i3}^*(\theta) \tag{6.3}
\]

\[
P_{i3}(\theta) = P_{i3}^*(\theta) - 0 \tag{6.4}
\]

\[
P_{ix}(\theta) = P_{ix}^*(\theta) - P_{i(x+1)}^*(\theta) \tag{7}
\]

### 1.4.3 Continuation Ratio Models

A third class of models suited for the analyses of polytomous data is the class of continuation ratio models (CRMs) (Mellenbergh, 1995; Hemker et al., 2001). Polytomous items with a sequential scoring mechanism determining the response outcome, are especially suited for this class of models (Agresti, 2013; Hemker et al., 2001) and they are referred to as sequential models (SMs; Tutz, 1990). Akkermans (2000) clarified a sequential scoring rule, based on an example mathematics item of Masters (1982). The example is as follows: \( \sqrt{7.5/0.3 - 16} = ? \) In order to answer this item correctly, three calculations have to be performed. These calculations are 1) \( 7.5/0.3 = 25 \); 2) \( 25 - 16 = 9 \); and 3) \( \sqrt{9} = 3 \). If the item is scored sequentially, one point is given when the first step is correctly solved, two points are given when the first two steps are correctly solved and three points are given when, in addition to the first two steps, also the last step is carried out correctly. An item step is conceptually a dichotomous Rasch item (Verhelst, Glas & de Vries, 1997) and a subject is only administered
the next, in concept Rasch item, if a correct response was given to the previous one. So it is assumed that a subject keeps taking item steps until an incorrect response is given (Verhelst et al., 1997).

The response categories of this four-category example are \( x = 0, 1, 2, 3 \). In the CRM, the probabilities of \( x = 1 \) and higher vs. \( 0 \), \( x = 2 \) and higher vs. \( 1 \) and \( x = 3 \) vs. \( 2 \) are calculated (De Boeck & Partchev, 2012). So the ordinal nature of the response variable is preserved by splitting the \( k \) categories into \((k - 1)\) continuation ratios (Mellenbergh, 1995), see Figure 4.

<table>
<thead>
<tr>
<th>Four ordered categories</th>
<th>Continuation Ratios</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Categories</td>
</tr>
<tr>
<td></td>
<td>1, 2 and 3 vs. 0</td>
</tr>
<tr>
<td></td>
<td>Categories 2 and 3 vs. 1</td>
</tr>
<tr>
<td></td>
<td>Category 3 vs. 2</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>0</th>
<th>0</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1 &amp; 2 &amp; 3</td>
<td>2 &amp; 3</td>
</tr>
<tr>
<td>2</td>
<td></td>
<td>3</td>
</tr>
</tbody>
</table>

*Figure 4. Continuation Ratio Model.*

The conditional probability of passing an item step is given by the following equation:

\[
M_{ix}(\theta) = \frac{P(X_i \geq x|\theta)}{P(X_i \geq x - 1|\theta)}
\]

(8)

where when \( x = 0 \), \( M_{io}(\theta) \) equals 1 for all \( \theta \).

\( M_{ix}(\theta) \) is called the *item step response function* (ISRF). In this formula, \( M_{i1}(\theta) \) is calculated from the probability of a response to fall in or above category 1 (so 1 through 3), divided by the probability of a response to fall in or above category 0 (0 through 3). These probabilities can be calculated with the use of several models (e.g., acceleration model, 1-parameter sequential model and 2-parameter sequential model).

The conditional probability of responding in category \( x \) of item \( i \) is written as the product of the ISRFs for the \( x \) steps that were successfully solved and the conditional probability of failing step \( x + 1 \) given that the previous steps were successfully solved (Hemker et al., 2001). This conditional probability is written as follows:

\[
P(X_i = x|\theta) = \prod_{y=0}^{x} M_{iy}(\theta)[1 - M_{i,x+1}(\theta)]
\]

(9)
One type of model inspired from sequential models is the response tree model (De Boeck & Partchev, 2012). In response tree models, the response categories are represented with a binary response tree and the response process can be interpreted as a sequential process of going through the tree to its end nodes (De Boeck & Partchev, 2012). A response tree regarding the four-category example from Masters (1982) is presented in Figure 5 whereby $X^*$ present internal nodes and $x$ the response categories. Figure 5 shows a linear response tree since one branch from each internal node ($X^*$) directly leads to an end node, i.e., a response category (De Boeck & Partchev, 2012).

![Figure 5. Linear response tree for the four response categories.](image)

It can be seen that from the top node ($X_1^*$), the left branch leads directly to an end node (response category 0) while the right branch leads to the second internal node. The first internal node is called sub-item 1, with the left branch coded as 0 and the right branch coded as 1. The second internal node ($X_2^*$) is then called sub-item 2 with the left branch again coded as 0 and the right as 1. This is also the case for the last internal node, which is sub-item 3. So, the non-analogical other strategy (0) is recoded in terms of the sub-items as (0, NA, NA) because the first sub-item score is 0 and the others sub-items scores are not applicable (NA). For all four response categories this will lead to the following mapping matrix $T$ presented in Figure 6.
It is clear that the original item responses are denoted as $x$ ($x = 0, 1, 2, 3$). The sub-item responses $X_{ij}^*$ are denoted as NA, 0, or 1 with $j$ ($r = 1, \ldots, J$) as index for the sub-items, one per node (De Boeck & Partchev, 2012). For an item with four response categories, assuming one underlying latent trait variable for all nodes, the probabilities of answering in a certain response category are presented in the following equations (9.1 to 9.4):

$$
\pi(X_i = 0|\theta) = \pi(X_{i1}^* = 0|\theta) \tag{9.1}
$$

$$
\pi(X_i = 1|\theta) = \pi(X_{i1}^* = 1|\theta) \pi(X_{i2}^* = 0|\theta) \tag{9.2}
$$

$$
\pi(X_i = 2|\theta) = \pi(X_{i1}^* = 1|\theta) \pi(X_{i2}^* = 1|\theta) \pi(X_{i3}^* = 0|\theta) \tag{9.3}
$$

$$
\pi(X_i = 3|\theta) = \pi(X_{i1}^* = 1|\theta) \pi(X_{i2}^* = 1|\theta) \pi(X_{i3}^* = 1|\theta) \tag{9.4}
$$

The probabilities of the left and right branches from each node $\pi(X_{ij}^* = 0, 1)$ are determined by a logistic regression model (De Boeck & Partchev, 2012). This is presented in the following equation 10:

$$
\pi(X_{ij}^* = x_{ij}^*|\theta) = \frac{\exp(\theta + \beta_{ij})^{x_{ij}}}{1 + \exp(\theta + \beta_{ij})} \tag{10}
$$

where $\theta$ is the subject’s latent trait level and $\beta_{ij}$ the minus item difficulty or the threshold.

### 1.5 Appropriate method for the study of repeated measures data

Educational and psychological research is often interested in the change of trait level over time or after a certain treatment of training. In this study we specifically examine the change in children’s strategy use; thus we are dealing with repeated measures data.

Embretson (1991) proposed a multidimensional Rasch model for the measurement of learning and change (MRMLC) based on item response theory. In this model, it is assumed that on the first measurement occasion ($k = 1$), performance depends on the initial ability of a person. In addition, at subsequent measurement occasions ($k > 1$), performance depends on initial ability as well as ($k - 1$) additional abilities which are called modifiabilities (Embretson, 1991;
von Davier, Xu & Carstensen, 2011). The MRMLC gives the probability that a subject $s$ passes item $i$ on occasion $k$ as follows:

$$
P(X_{isk} = 1|\theta_{si}, ..., \theta_{sk}, \beta_i) = \frac{\exp(\sum_{m=1}^{k}\theta_{sm} - \beta_i)}{1 + \exp(\sum_{m=1}^{k}\theta_{sm} - \beta_i)}$$

where $\theta_i$ is a vector of abilities so $\theta_{i1}$ represents the initial ability at the first measurement occasion $k = 1$ and the modifiabilities ($\theta_{sm}$ with $m > 1$) represent the additional abilities from previous measurement occasions. This model shows that for item $i$ on occasion $k$, all abilities up to occasion $k$ are involved (Embretson, 1991). So across conditions, the MRMLC is a multidimensional model.

Stevenson, Hickendorff, Resing, Heiser and de Boeck (2013) applied the MRMLC with an extension of explanatory variables in order to measure initial analogical reasoning ability and performance change after training. They dynamically tested analogical reasoning skills of 252 children using a pretest-training-posttest design. Two training conditions were applied; graduated prompts and outcome feedback. The graduated prompts training consisted of stepwise instructions in order to help the child solve the analogy problem. In the outcome feedback training a child was only told whether the given answer was correct or incorrect.

In addition to a simple IRT model with random intercepts for both persons and items, Stevenson et al. (2013) included a fixed session parameter to model the average change from pretest to posttest and random session parameters to allow the session effect to vary over persons. After fitting this model to the data, they concluded that there were individual differences in the change from pretest to posttest regarding analogical reasoning skills. Children trained with graduated prompts improved more than children who received the outcome feedback training. In addition, children who scored lower at the pretest tended to improve more after training than children with higher pretest scores.

In this study, we will also apply Embretson’s vision regarding the way the latent abilities are related to the different measurement occasions. The model will be generalized to a polytomous IRT model but will use the same basics of an initial trait level and modifiability. Thus, at time of the pretest, only the initial ability level will be involved in performance. At time of the posttest, initial ability plus an additional ability will be involved.
2. Research questions

In this study two research questions are addressed. The first research question is substantive in nature and aimed at gaining more insight in the use of strategies in solving analogical reasoning tasks: ‘Are there training- and age related changes regarding the strategies children use to solve figural analogical problems?’ Expected is that children trained with the a more comprehensive training will on average improve more in analogy solving than children trained with a less comprehensive training and will therefore use analogical correct strategies more often (Stevenson et al., 2013). In addition, older children are expected to be better in solving analogy problems and younger children to generally improve more (e.g., Hosenfeld et al., 1997). In line with this expectation, younger children are expected to generally benefit more from the more comprehensive training conditions compared to older children.

The second question of this research is methodological and concerns the models used to investigate this type of data. In order to derive answers to the above research question, the three previously discussed appropriate polytomous IRT models will be fitted to the data. The research question is formulated as follows: ‘Which polytomous IRT model (PCM, GRM or CRM) is most appropriate for the analyses of the current data?’ This answer will be based on several important guidelines in model selection. As addressed in the introduction, the three models differ theoretically from each other. However, based on previous research (Nering & Ostini, 2010), it is expected that the three different polytomous IRT models fitted on the current data set will not lead to substantially different measurement outcomes.
3. Method

3.1 Design and procedure

Over four years, six analogical reasoning experiments were conducted at different schools and in different grades. The experiments are named after the year of administering, resulting in experiments 20091, 20092, 20101, 20102, 20111 and 20121. Each experiment had a pretest-training-posttest design. All participating children were paired based on age, gender, classroom and cognitive ability estimates and then randomly assigned to different training conditions. In total, there were four types of training conditions.

Each session (pretest, training and posttest) was conducted within approximately 20 minutes, individually in a quiet room at the participant’s school and by a trained psychology student. On average, the posttest was administered two weeks after the pretest.

3.2 Participants

A total of 1033 school children participated in the study. The children were recruited from different elementary schools in the Netherlands. Schools were selected based on their willingness to participate. From the parents, a written informed consent was obtained prior to participation. After excluding 26 children (teacher not willing to participate, child moving to different school, no permission obtained by parents), the total sample contained the responses of 1007 children. Approximately as many boys as girls (490 boys and 517 girls) were enrolled with a mean age of 7.3 years (90% range 5.2 – 10.2). Each experiment was conducted in a different grade (or in multiple grades) and with different participants (see Table 1).
Table 1

*Characteristics per Experiment*

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>20091</th>
<th>20092</th>
<th>20101</th>
<th>20102</th>
<th>20111</th>
<th>20121</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male</td>
<td>75 (51.7)</td>
<td>42 (60.9)</td>
<td>25 (49.0)</td>
<td>115 (44.6)</td>
<td>117 (46.4)</td>
<td>116 (50.0)</td>
<td>490 (48.7)</td>
</tr>
<tr>
<td>Age</td>
<td>5.5 (0.3)</td>
<td>5.3 (0.3)</td>
<td>7.1 (0.6)</td>
<td>7.0 (0.4)</td>
<td>7.0 (1.0)</td>
<td>9.6 (0.7)</td>
<td>7.3 (1.6)</td>
</tr>
<tr>
<td>Grade</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>17 (24.6)</td>
<td>0</td>
<td>0</td>
<td>5 (2.0)</td>
<td>0</td>
<td>22 (2.2)</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>145 (100.0)</td>
<td>52 (75.4)</td>
<td>0</td>
<td>0</td>
<td>70 (27.8)</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>0</td>
<td>24 (47.1)</td>
<td>258 (100.0)</td>
<td>90 (35.7)</td>
<td>0</td>
<td>372 (36.9)</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>0</td>
<td>27 (52.9)</td>
<td>0</td>
<td>87 (34.5)</td>
<td>0</td>
<td>114 (11.3)</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>99 (42.7)</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>6</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>133 (57.3)</td>
<td>0</td>
</tr>
<tr>
<td>Training type</td>
<td>C</td>
<td>0</td>
<td>69 (100.0)</td>
<td>26 (51.0)</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>P</td>
<td>70 (48.3)</td>
<td>0</td>
<td>0</td>
<td>131 (50.8)</td>
<td>0</td>
<td>78 (33.6)</td>
</tr>
<tr>
<td></td>
<td>OF</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>125 (49.6)</td>
<td>77 (33.2)</td>
</tr>
<tr>
<td></td>
<td>GP</td>
<td>75 (51.7)</td>
<td>0</td>
<td>25 (49.0)</td>
<td>127 (49.2)</td>
<td>127 (50.4)</td>
<td>77 (33.2)</td>
</tr>
</tbody>
</table>

1Values are n (%).
2Values are mean (SD).
3C = control, P = practice, OF = outcome feedback, GP = graduated prompts.

3.3 Material

3.3.1 Figural analogy task

In order to assess analogical reasoning, a computerized dynamic test called AnimaLogica was used (Stevenson et al., 2013). In this test several figural analogies were presented consisting of a 2x2 matrix (see Figure 7) with familiar animals. In order to get the right picture in the empty box (A:B::C:D), participants had to construct the solution using a computer mouse and drag and drop the animal figures to this box (an example item is presented in Figure 9). The empty box was either in the lower left or the lower right quadrant of the matrix. Within each figural analogy, horizontal as well as vertical transformations were possible resulting in one total number of transformations. The transformations possible were type of animal (camel, bear, dog, horse, lion and elephant), color (yellow, blue and red), orientation, position, quantity (one or two) and size (small and large). For example, two horizontal transformations apply to the figural analogy presented in Figure 7 namely size and position. Vertically, three transformations apply (animal type, orientation and quantity) which results in the total number of transformations to be five. The number of transformations was related to
the difficulty of each item. Within the current experiments, item difficulty ranged from two to eight total transformations. This can be seen by the first number of every item (itemcode) in Figure 8, which represents the number of transformations.

![Figural analogy from AnimaLogica.](image)

**Figure 7.** Figural analogy from AnimaLogica.

### 3.3.2 Pre- and posttest

A pretest-training-posttest design focuses on measuring the change (the potential for learning) in participants analogical reasoning skills brought about by training. The pretest provides an indication of the participant’s initial ability regarding analogical reasoning (Stevenson, 2012). After training (that will be discussed in the next paragraph), the posttest provides information about the potential for learning, in other words the potential ability (Stevenson, 2012).

The items administrated in the pretest and posttest were isomorphs, meaning that they could differ in color and animal type but had to be solved using the same transformations. Therefore their difficulties are assumed to be equal. Exceptions were items 605 and 710 in experiment 20091 and 20092 and items 401 and 511 in experiment 20111 (see Figure 8). In these cases, the items accidentally different in one transformation resulting in a slightly different item. The number of items and the items themselves varied between experiments. The experiments contained 15, 15, 18, 20, 20 and 24 items respectively. Within the different experiments, there were a number of overlapping items. Seven items were included in all experiments (201, 204, 301, 404, 502, 505 and 604). The total number of administered items was 35.
### Figure 8

Items administrated per experiment. Dark grey indicates a pretest-item, light grey indicates a posttest-item.
3.3.3 Training

In total, there were four different training conditions. Table 1 shows which training type was used in which experiment. The most comprehensive training condition was the graduated prompts technique. With this technique, as explained in Stevenson et al. (2013), stepwise instructions were given to the participant starting with general, metacognitive prompts and ending with step-by-step scaffolds to solve the problem. When a participant solved a problem correctly, he or she was asked to explain the given answer after which no more prompts were given. The total number of graduated prompts given ranged between zero (when the problem was solved correctly in the first attempt) and five.

The second most comprehensive training condition was outcome feedback. As explained in Stevenson et al. (2013), outcome feedback training allows the participant to have four attempts in order to correctly solve a problem. With each attempt, the participant was told if their answer was correct or incorrect and they received motivational comments. After four attempts, regardless whether the problem was solved correctly or not, the participant proceeded to the next training item.

The third training condition was practice without feedback. Hereby participants were presented with the same training items as the other training conditions, except they did not receive any feedback.

The last training condition classified was the control group in which children did not practice with figural analogies at all. In each experiment, two of the four training conditions were assigned to the participants.

3.4 Variables

3.4.1 Response variable

The recorded response was the strategy used for the solution of the analogical problem. This response was directly derived from the participants answer to the analogical problem. The strategies were classified into four main categories, namely 1) correct analogical; 2) partial analogical; 3) duplication non-analogical; and 4) other non-analogical. An example of each solution strategy is presented in Figure 9. Correct analogical was recorded when the item was answered correctly. Partial analogical was recorded when the answer was missing one or two transformations. When the duplication non-analogical strategy was applied, a participant had copied one of the already visible matrix quadrants. Other non-analogical was recorded when three or more transformations were missing.
Strategy use is an ordinal variable. The highest level of performance on each item is the correct analogical strategy. In decreasing level of performance, correct analogical is followed by partial analogical, duplicate non-analogical and other non-analogical strategies resulting in an ordinal variable. The child’s recorded strategy on a particular item will be the dependent variable.

3.4.2 Person predictors

To be able to answer what the effect of training and age are on the analogical reasoning skills, an explanatory IRT model is necessary. With explanatory models, the item responses are explained in terms of other variables (De Boeck & Wilson, 2004) by estimating the effects of predictor variables on the latent factor(s) (Hickendorff, 2013). These predictors can be on
person level, item level or on person-by-item level (De Boeck & Wilson, 2004). In this study, these variables are on person level and are therefore called person predictors. Since we are interested in the effects of training condition and age on the strategies in analogical reasoning, these person predictors will be included in the model resulting in a latent regression analysis (Hickendorff, 2013) in which the latent traits are considered to be regressed on the external person predictors (De Boeck & Wilson, 2004). The predictor variable age will be centered around its mean by subtracting the mean age from each observed age. This way, the meaning of the intercept changes. When the value of the predictor age is 0, the intercept value represents the analogical reasoning skills of a child with average age instead of a 0-year-old.

As previously presented, there are four different training conditions, namely graduated prompts, outcome feedback, practice and control. This person predictor is dummy-coded in 3 binary predictors with graduated prompts as reference category. The person predictor age is a continuous variable and will be reported as age in months.

3.5 Structural model

The measurement models that will be applied in this study were presented in Section 1. These are the partial credit model, graded response model and the continuation ratio model. Here, the structural model will be discussed.

The structural model is presented in Figure 10. Observed variables are represented by rectangles and the latent variables by circles. Arrows represent regressions. The dotted line framing the structural model represents the six experiments. As mentioned earlier, responses of all participants are administered of both the pretest and the posttest. This type of data is often characterized by response dependencies within persons – that is, within-subject correlation (De Boeck & Wilson, 2004). To be able to incorporate the correlation of the trait levels across both test occasions, a multidimensional approach is necessary. In other words, following Embretson’s approach, the latent structure is regarded to be multidimensional with the first dimension to be the trait level at time of the pretest (represented by $\theta_0$) and the second dimension the modifiability (represented by $\theta_1$). The modifiability refers to the performance change from pretest to posttest. This is presented in Figure 10. Figure 10 also shows that the person property age will be added to both dimensions. Training condition will only be added to the second dimension (posttest trait level) since it has no influence on the initial ability.
3.6 Trait level distributions per experiment

In this current study the responses of children of six different experiments are used. Responses of children that belong to the same experiment (cluster) can be expected to be more similar than the responses of children who belong to different experiments. In this case, it cannot be assumed that all children are sampled from the same common distribution. Therefore, we will investigate this experiment effect by assuming separate distributions for children belonging to different experiments. Thus, the latent variable of a subject $\theta_g$ depends on group $g$ with $\theta_g \sim N(\mu_g, \sigma_g^2)$.

3.7 Model selection

One aim of this study is to find the most appropriate polytomous IRT model, from a selection of three, for the analyses of the current data. Of course, the most appropriate model can be defined and therefore interpreted in different ways. If the goal is to find the model with maximum fit to a certain data set, the model with the smallest root mean squared deviation between the observed and the expected responses may be the best and therefore most appropriate model (Sung & Kang, 2006). However, on the other hand, the goal can also be to
find the model with the clearest interpretation and answer to the research question. In this study, the most appropriate model will be based on the model fit in association with parsimony, the interpretation of the parameters and substantial features of the data. In other words, the model that can explain the important features of the data without adding unnecessary complexity.

To evaluate and compare model fit, three fit indices will be reported for each model. These indices are the deviance, the Akaike information criterion (AIC) (Akaike, 1974) and the Bayesian information criterion (Schwarz, 1978). They can be used to compare non-nested models, which is the case in this study. The deviance is defined as $-2\log(L_M - L_S)$ with $L_M$ as the maximized log-likelihood value for a model M of interest, and $L_S$ as the maximized log-likelihood value for the saturated model (Agresti, 2013). The saturated model is the most complex model with a parameter for every observation so that it provides a perfect fit to the data. Thus, the deviance is the likelihood ratio statistic for comparing a model of interest to the saturated model. The AIC and the BIC are derived from the deviance and are penalized-likelihood criteria with a penalty included for the number of parameters. The number of parameters is of course very important to take into account when evaluating a model. The AIC and BIC are usually written as $-2\log(L) + kp$ where $L$ is the likelihood function, $p$ is the number of parameters and $k$ is 2 for the AIC and $\log(n)$ for the BIC with $n$ being the number of persons. The lower the value of these three indices, the better the fit of the model (De Boeck & Wilson, 2004).

Since both the PCM and the GRM are estimated with the same software package mirt, their AIC and BIC can directly be compared with each other. For the CRM, it should be taken into account that the lme4 package employs different estimation methods and model parameterization.

### 3.8 Statistical analyses

Analyses will be performed using SPSS (version 22) and R Statistical Software (R Development Core Team, 2013).

#### 3.8.1 Software

The previously discussed PCM and GRM will be fitted on the data using the R-package mirt, which stands for multidimensional item response theory (Chalmers, 2012). mirt provides uni- and multidimensional latent trait models under the Item Response Theory paradigm for binary and polytomous item responses. It contains many flexible parameter
estimation features. mirt fits an unconditional maximum likelihood factor analysis model using either the MHRM (Metropolis-Hastings Robbins-Monro) algorithm developed by Cai (2010) or with an EM (expectation-maximization) algorithm (outlined by i.e., De Boeck & Wilson, 2004) using rectangular or quasi-Monte Carlo integration grids (Chalmers, 2012).

The CRM will be fitted using the lme4 package in R (Bates, Mächler, Bolker & Walker, 2014). This R-package provides functions for fitting and analyzing linear mixed models, generalized linear mixed models and nonlinear mixed models (Bates, 2014). The default estimation method that the lme4 package uses is the Laplace approximation of the likelihood (Bates et al., 2014).
4. Results

4.1 Psychometric properties

In order to determine the internal consistency of the pre- and posttest, Cronbach’s alpha coefficients were calculated per experiment. The Cronbach’s alpha is a method for the estimation of reliability (Furr & Bacharach, 2008). Cronbach’s alphas for the six pretests were $\alpha(20091) = .74$, $\alpha(20092) = .78$, $\alpha(20101) = .91$, $\alpha(20102) = .89$, $\alpha(20111) = .93$ and $\alpha(20121) = .90$. For the six posttests these were $\alpha(20091) = .89$, $\alpha(20092) = .86$, $\alpha(20101) = .94$, $\alpha(20102) = .93$, $\alpha(20111) = .94$ and $\alpha(20121) = .88$. All Cronbach’s alphas indicate good to excellent internal consistencies of the tests.

4.2 Proportion of strategy use

As mentioned earlier, seven items were included in all experiments. Table 2 presents the proportion of strategy use per item on both test occasions (pretest and posttest). The pretest proportion of analogical correct use of strategy per item ranged from .09 to .41 and for the posttest from .26 to .61. The Spearman rank correlation between this proportion and the predicted difficulty level based on the number of transformations was $\rho = -.982$, $p < .001$ for the pretest and $\rho = -.982$, $p < .001$ for the posttest. These strong correlations indicate that as the number of transformations increased, the proportion of analogical correct strategy use decreased. So, the number of transformations is a good predictor of item difficulty. This can also be seen in Table 2 knowing that the first digit of the item represents the number of transformations. For example, items 201 and 204 with each 2 transformations, have been solved more often with an analogical correct strategy than items 502 and 505 with each 5 transformations.

Another observation that can be made from Table 2 is that for each item, the proportion of non-analogical other strategies decreased from pretest to posttest. This also applies to the non-analogical duplication strategy. For the analogical partially correct strategy it is less clear to define a trend. For item 201, 202, 301 and 404 the proportion decreased from pretest to posttest. For item 502, it remains the same and for 505 and 604 the proportion of this strategy increased. Finally, we can see that items were more often solved with an analogical correct strategy in the posttest than in the pretest.
Table 2
Proportion of Strategy Use per Common Item

<table>
<thead>
<tr>
<th>Item</th>
<th>Test occasion</th>
<th>Non-analogical (other)</th>
<th>Non-analogical (duplication)</th>
<th>Analogical (partially correct)</th>
<th>Analogical (correct)</th>
<th>N</th>
</tr>
</thead>
<tbody>
<tr>
<td>201</td>
<td>Pretest</td>
<td>0.05</td>
<td>0.26</td>
<td>0.32</td>
<td>0.37</td>
<td>1002</td>
</tr>
<tr>
<td></td>
<td>Posttest</td>
<td>0.03</td>
<td>0.14</td>
<td>0.25</td>
<td>0.57</td>
<td>992</td>
</tr>
<tr>
<td>204</td>
<td>Pretest</td>
<td>0.04</td>
<td>0.35</td>
<td>0.20</td>
<td>0.41</td>
<td>1002</td>
</tr>
<tr>
<td></td>
<td>Posttest</td>
<td>0.01</td>
<td>0.25</td>
<td>0.13</td>
<td>0.61</td>
<td>992</td>
</tr>
<tr>
<td>301</td>
<td>Pretest</td>
<td>0.06</td>
<td>0.25</td>
<td>0.35</td>
<td>0.34</td>
<td>1002</td>
</tr>
<tr>
<td></td>
<td>Posttest</td>
<td>0.03</td>
<td>0.17</td>
<td>0.30</td>
<td>0.50</td>
<td>992</td>
</tr>
<tr>
<td>404</td>
<td>Pretest</td>
<td>0.19</td>
<td>0.23</td>
<td>0.39</td>
<td>0.19</td>
<td>1002</td>
</tr>
<tr>
<td></td>
<td>Posttest</td>
<td>0.10</td>
<td>0.14</td>
<td>0.37</td>
<td>0.39</td>
<td>992</td>
</tr>
<tr>
<td>502</td>
<td>Pretest</td>
<td>0.35</td>
<td>0.23</td>
<td>0.31</td>
<td>0.11</td>
<td>1002</td>
</tr>
<tr>
<td></td>
<td>Posttest</td>
<td>0.22</td>
<td>0.13</td>
<td>0.31</td>
<td>0.33</td>
<td>992</td>
</tr>
<tr>
<td>505</td>
<td>Pretest</td>
<td>0.34</td>
<td>0.25</td>
<td>0.30</td>
<td>0.11</td>
<td>1002</td>
</tr>
<tr>
<td></td>
<td>Posttest</td>
<td>0.20</td>
<td>0.14</td>
<td>0.31</td>
<td>0.35</td>
<td>992</td>
</tr>
<tr>
<td>604</td>
<td>Pretest</td>
<td>0.39</td>
<td>0.21</td>
<td>0.30</td>
<td>0.09</td>
<td>1002</td>
</tr>
<tr>
<td></td>
<td>Posttest</td>
<td>0.28</td>
<td>0.13</td>
<td>0.33</td>
<td>0.26</td>
<td>991</td>
</tr>
</tbody>
</table>

4.3 Methodological research question: ‘Which polytomous IRT model (PCM, GRM or CRM) is most appropriate for the analyses of the current data?’

To be able to answer the methodological research question, the partial credit model (PCM), graded response model (GRM) and continuation ratio model (CRM) were fitted to the common pretest items without any predictor effects. Hereby, all children were assumed to come from a single population with the same ability distribution. In addition, a second analyses was conducted under a multiple-group assumption since the current data is the aggregation of the data of six different experiments. Under the multiple-group assumption, children from different experiments were assumed to come from different populations with potentially different ability distributions (Von Davier, Xu & Carstensen, 2009).

The decision to only use the pretest items in order to answer this question was made so that there would be as little statistical noise as possible. During the pretest, there were less additional factors that could have influenced the analogical reasoning skills of the subjects which the figural analogy test aimed to measure. This way, the models can be properly compared to each other.
In addition, since the experiments included not only common but also unique items, we wanted to keep their comparability by only using the seven common items in the analyses corresponding to the methodological research question.

4.3.1 Partial Credit Model

The R-package mirt was used to fit the PCM to the common pretest items. The R-code of this model is presented in Appendix 1.1.1. The estimated parameters that mirt initially returns are the slope intercept parameters \(d\) with corresponding standard errors. These estimates are presented in Table A1 in Appendix 1.1.2. However, in terms of interpretation, they must not be mistaken with the traditional IRT parameters. Therefore, they were converted into the IRT parameters (in the case of the PCM into intersection parameters) and are presented in Table 3. The intersection parameters \(\delta_{i1}, \delta_{i2}\) and \(\delta_{i3}\) represent the relative difficulty of each step and are the points on the latent trait scale where two sequential category response curves intersect. To illustrate this, the category responses curves of item 204 are presented in Figure 11. At \(\delta_{i1} = -3.41\), a subject becomes relatively more likely to respond in category 1 than in category 0. In addition, at a trait level of 0.07, a subject becomes relatively more likely to respond in category 3 than 1 given that the subject already reached the first category. Note that for item 204, category 2 is never the most likely option. This can also be seen from the unordered intersection parameters of item 204 in Table 3. When the intersection parameters are unordered it means that, conditional on trait level, at least one category is never the most likely option.

Similar to the conclusions from Table 2, Table 3 shows the intersection parameters to gradually shift upwards on the latent trait continuum from item 201 to item 604. If we look at the intersection parameters \(\delta_{i3}\) from item 204 to 604, it is clear that the value increases with each item. Thus, subjects need increasingly higher trait levels in order to become more likely to respond in category 3, which is using a correct analogical strategy to solve a figural analogy task.
Table 3

<table>
<thead>
<tr>
<th>Item</th>
<th>$\delta_1$</th>
<th>$\delta_2$</th>
<th>$\delta_3$</th>
</tr>
</thead>
<tbody>
<tr>
<td>201</td>
<td>-2.74</td>
<td>-0.68</td>
<td>0.19</td>
</tr>
<tr>
<td>204</td>
<td>-3.41</td>
<td>0.07</td>
<td>-0.35</td>
</tr>
<tr>
<td>301</td>
<td>-2.58</td>
<td>-0.82</td>
<td>0.40</td>
</tr>
<tr>
<td>404</td>
<td>-1.00</td>
<td>-0.69</td>
<td>1.48</td>
</tr>
<tr>
<td>502</td>
<td>-0.14</td>
<td>-0.14</td>
<td>2.15</td>
</tr>
<tr>
<td>505</td>
<td>-0.25</td>
<td>0.00</td>
<td>2.09</td>
</tr>
<tr>
<td>604</td>
<td>0.10</td>
<td>-0.12</td>
<td>2.42</td>
</tr>
</tbody>
</table>

Figure 11. Category Response Curves of item 204 under the PCM.

4.3.2 Graded Response Model

The GRM was also fitted using mirt and the R-code is presented in Appendix 1.2.1. Appendix 1.2.2 shows the estimated slope intercepts parameters and their standard errors. These parameters are converted into between category threshold parameters ($\beta_j$'s) for the interpretation of the GRM and presented in Table 4. They represent the point on the latent trait scale where a subject had a probability of .50 of responding in or above category $j = x$. For example, for item 301, a subject with a trait level of -3.10 had a .50 probability of responding in or above category 1. With a trait level of -0.98, this subject had a .50 probability of responding in or above category 2 etc. This can also be seen in Figure 11, which presents the category response curves of item 301. The between category threshold parameters are ordered
within each item, which contrary to the PCM must occur in the GRM (Embretson & Reise, 2000).

Since the GRM is a 2PL model, the item slope parameters \( (\alpha_i) \) were estimated. Generally, the value of the item slope parameter represents the amount of information provided by the item. For example, item 502 has the largest slope parameter which leads to more peaked category response curves as can be seen in Figure 12. This indicates that the item is very capable to distinguish locally between subjects with different trait levels.

<table>
<thead>
<tr>
<th>Item</th>
<th>( \alpha )</th>
<th>( \beta_1 )</th>
<th>( \beta_2 )</th>
<th>( \beta_3 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>201</td>
<td>1.78</td>
<td>-2.19</td>
<td>-0.70</td>
<td>0.40</td>
</tr>
<tr>
<td>204</td>
<td>1.51</td>
<td>-2.73</td>
<td>-0.45</td>
<td>0.31</td>
</tr>
<tr>
<td>301</td>
<td>1.02</td>
<td>-3.10</td>
<td>-0.98</td>
<td>0.74</td>
</tr>
<tr>
<td>404</td>
<td>2.05</td>
<td>-1.11</td>
<td>-0.29</td>
<td>1.20</td>
</tr>
<tr>
<td>502</td>
<td>3.13</td>
<td>-0.51</td>
<td>0.14</td>
<td>1.48</td>
</tr>
<tr>
<td>505</td>
<td>2.24</td>
<td>-0.56</td>
<td>0.21</td>
<td>1.60</td>
</tr>
<tr>
<td>604</td>
<td>1.94</td>
<td>-0.45</td>
<td>0.26</td>
<td>1.86</td>
</tr>
</tbody>
</table>

*Table 4 Estimated Item Parameters of the Graded Response Model*

*Figure 12. Category Response Curves of item 502 under the GRM.*

### 4.3.3 Continuation Ratio Model

As previously mentioned, one class of continuation ratio models are the so-called sequential models in which a sequential process determines the response outcome. A response tree model, inspired from these sequential models, was fitted to the data next. In order to fit a response tree model using the function `glmer` from the `lme4` package, the data had to be transformed into the form required by `glmer`. Using the R-package `irtrees`, the mapping
matrix $T$ (as presented in Figure 6) was applied to the data such that each line of the data matrix pertains to a person and a sub-item (an item node) (De Boeck & Partchev, 2012). After preparation, a unidimensional model for linear response trees was fitted to the seven common pretest items, referred to as CRM (R-code is presented in Appendix 1.3). Items and nodes were included in the model as fixed effects.

The parameter estimates are presented in Table 5. They represent the propensity of scoring 1 instead of 0 at an internal branch (as presented in Figure 5). For all items, the easiness parameters of the second subitem ($\beta_2$) were lower than the easiness parameters of the first subitem ($\beta_1$). So, as the number of steps of an item increases, the probability of ending in a left branch and thus to make a mistake, also increases. This makes sense since the step to a higher quality response becomes more difficult since this requires better analogical reasoning skills. Figure 13 shows the category response curves of item 505. As can be seen, with a higher trait level, the probability of using a correct analogical strategy to solve the item becomes more likely.

Comparable to the output of the PCM and GRM, it can be seen from Table 5 that the items become more difficult in a relative consecutive order.

<table>
<thead>
<tr>
<th>Item</th>
<th>$\beta_1$ (SE)</th>
<th>$\beta_2$ (SE)</th>
<th>$\beta_3$ (SE)</th>
</tr>
</thead>
<tbody>
<tr>
<td>201</td>
<td>3.79 (0.16)</td>
<td>1.29 (0.10)</td>
<td>-0.23 (0.11)</td>
</tr>
<tr>
<td>204</td>
<td>4.25 (0.19)</td>
<td>0.78 (0.09)</td>
<td>0.37 (0.11)</td>
</tr>
<tr>
<td>301</td>
<td>3.68 (0.16)</td>
<td>1.39 (0.10)</td>
<td>-0.39 (0.11)</td>
</tr>
<tr>
<td>404</td>
<td>2.02 (0.11)</td>
<td>1.00 (0.10)</td>
<td>-1.83 (0.12)</td>
</tr>
<tr>
<td>502</td>
<td>0.93 (0.09)</td>
<td>0.30 (0.11)</td>
<td>-2.66 (0.14)</td>
</tr>
<tr>
<td>505</td>
<td>0.98 (0.09)</td>
<td>0.22 (0.11)</td>
<td>-2.57 (0.14)</td>
</tr>
<tr>
<td>604</td>
<td>0.69 (0.09)</td>
<td>0.39 (0.11)</td>
<td>-2.91 (0.15)</td>
</tr>
</tbody>
</table>
4.3.4 Multiple groups

The current data is an aggregation of the data of the six experiments. Therefore, there is a certain clustering present in the data; children from different experiments were assumed to come from different populations with potentially different ability distributions. Therefore, the latent variable of a subject \( \theta_g \) depends on group \( g \) and \( \theta_g \sim N(\mu_g, \sigma_g^2) \). A factor variable indicating group membership (experiment) was incorporated in all three models (PCM, GRM and CRM) resulting in the models PCM2, GRM2 and CRM2.

For the PCM2 and GRM2 fitted with mirt, it was sufficient to perform a full-information maximum-likelihood multiple group analysis using the option multipleGroup (Chalmers, 2012). The R-code for the analyses of these models is presented in Appendix 2.1 and 2.2. Experiment 20091 was set as reference group. For this reference group, \( \mu_g \) is set to zero and \( \sigma_g \) to 1. For the other groups the mean and variance were estimated freely. The estimated item parameters were constrained to be equal across groups and therefore possible differential item functioning was ignored.

For the CRM fitted using lme4, the multiple groups aspect was added to the model both as random and fixed effects, resulting in CRM2 (the R-code is presented in Appendix 2.3). A random-effect term was added for each experiment, since there might be some variability in test scores due to different experiments (Bates, Mächler, Bolker & Walker, in press). With a random effect-term for each experiment, experiments are allowed to have random intercepts and slopes. The interpretation of the experiment’s fixed-effect terms is that these are the estimated population mean values of the random intercept and slope (Bates et al., in press).
The estimated latent means of the six experiments were relatively comparable regarding the three models. They are presented in Table 6. Experiment 20121 had the highest estimated latent mean compared to experiment 20091. This makes sense, since the children from experiment 20121 were older (grade 5 and 6) than the children from 20091 (grade 2; Table 1). Since older children are better at solving analogical problems than younger children, the mean theta would logically be higher in experiments with older children compared to experiments with younger children.

The within-experiment variances (Var) are also reported in Table 6. As can be seen, the random effects of the experiments in model CRM2 are equal indicating that the variability of test scores within the experiments are similar. Table 6 shows that for the PCM2 and GRM2 the variances within each experiment are not equal for all experiments.

<table>
<thead>
<tr>
<th>Experiment</th>
<th>PCM2 M</th>
<th>PCM2 Var</th>
<th>GRM2 M</th>
<th>GRM2 Var</th>
<th>CRM2 M</th>
<th>CRM2 Var</th>
</tr>
</thead>
<tbody>
<tr>
<td>20091 (Reference)</td>
<td>0.00</td>
<td>0.34</td>
<td>0.00</td>
<td>1.00</td>
<td>0.00</td>
<td>1.30</td>
</tr>
<tr>
<td>20092</td>
<td>0.37</td>
<td>0.23</td>
<td>0.40</td>
<td>0.57</td>
<td>0.52</td>
<td>1.30</td>
</tr>
<tr>
<td>20101</td>
<td>1.19</td>
<td>0.25</td>
<td>1.64</td>
<td>0.38</td>
<td>1.57</td>
<td>1.30</td>
</tr>
<tr>
<td>20102</td>
<td>0.28</td>
<td>0.74</td>
<td>0.34</td>
<td>1.30</td>
<td>0.42</td>
<td>1.30</td>
</tr>
<tr>
<td>20111</td>
<td>0.33</td>
<td>1.08</td>
<td>0.46</td>
<td>1.81</td>
<td>0.48</td>
<td>1.30</td>
</tr>
<tr>
<td>20121</td>
<td>2.32</td>
<td>1.07</td>
<td>2.74</td>
<td>1.13</td>
<td>2.98</td>
<td>1.30</td>
</tr>
</tbody>
</table>

4.3.5 Model selection

As described in the method section of this study, the most appropriate model will be based on the fit indices (deviance, AIC and BIC), interpretation of the parameters and substantial features of the data. These three arguments will be discussed regarding the PCM, GRM and CRM. Additionally, some practical concerns will be addressed. Since the models that take into account group membership rely on different sample spaces than the models with the single group assumption, it is difficult to compare them with each other. Therefore, the main focus of the model selection will be on the models without group structure.

4.3.5.1 Fit indices

The first argument in model selection are the fit indices. Table 7 presents the deviance, AIC, BIC and number of parameters of all models. It is clear that all three fit indices are lowest...
regarding the GRM, both with and without multiple groups. Since the GRM includes an item specific slope parameter ($\alpha_i$), this model contains more parameters than the other models and is, thus, more complex than the PCM and CRM. With more parameters, there is a larger sample size necessary for a good estimation of the parameters. However, in this current study, the sample size is considered to be rather large for an experimental study. Also, even though the AIC and BIC penalize on the number of parameters in the model, they are still the lowest in the graded response model (Agresti, 2013). Thus, based on the deviance, AIC and BIC, the graded response model would be the most appropriate model for the analysis of the current data.

<table>
<thead>
<tr>
<th>Model</th>
<th>Deviance</th>
<th>AIC</th>
<th>BIC</th>
<th>#p</th>
</tr>
</thead>
<tbody>
<tr>
<td>PCM</td>
<td>15708.7</td>
<td>15752.7</td>
<td>15860.8</td>
<td>22</td>
</tr>
<tr>
<td>GRM</td>
<td>15533.0</td>
<td>15589.0</td>
<td>15726.6</td>
<td>28</td>
</tr>
<tr>
<td>CRM</td>
<td>15624.4</td>
<td>15668.4</td>
<td>15837.9</td>
<td>22</td>
</tr>
<tr>
<td>PCM2</td>
<td>15128.0</td>
<td>15192.0</td>
<td>15349.3</td>
<td>32</td>
</tr>
<tr>
<td>GRM2</td>
<td>14970.0</td>
<td>15046.0</td>
<td>15232.7</td>
<td>38</td>
</tr>
<tr>
<td>CRM2</td>
<td>15101.0</td>
<td>15165.0</td>
<td>15411.5</td>
<td>32</td>
</tr>
</tbody>
</table>

4.3.5.2 Interpretation of the parameters

The second argument in the decision for the most appropriate model involves the interpretation of the parameters. All estimated item parameters from each model must be interpreted differently (Sijtsma, 2001). In the PCM, these parameter estimates are referred to as the step difficulties or category intersections and denoted by $\delta_{ij}$. In the GRM, these are the between category threshold parameters ($\beta_{ij}$) and in the CRM they are denoted by $\beta_{ij}$ and represented the item easiness parameters. One aspect that can help with the interpretation are the item plots as presented in Figure 11, 12 and 13. All item plots of the PCM, GRM and CRM are presented in the Appendix (Section 3). They seem relatively comparable with each other and show nothing counterintuitive. As can be seen from these plots, the category response curves of the items analyzed with the GRM were influenced by the slope parameter. All slope parameters were larger than 1 resulting in more peaked category response curves compared to the item plots of the PCM and CRM.

Another aspect one could evaluate are the standard errors of the item parameters. The standard errors reflect how rapidly the data likelihood changes around the parameter value. The more rapidly the likelihood changes, the smaller the standard error (Embretson & Reise, 2000).
If a model reports extremely large standard errors this could be an indication that the item parameter is an unstable estimation. For the current models, all standard errors were small.

Regarding the interpretation of the parameters of the three models, there were no noteworthy differences. This is in line with previous research. As described in Nering and Ostini (2010), there has been little demonstrated evidence that different polytomous IRT models produce substantially different measurement outcomes when applied to the same data.

4.3.5.3 Features of the data

The current data is polytomously scored and assumed to be ordinal in nature. Therefore, the three proposed IRT models are in theory, based on these data characteristics, all appropriate for the analyses of the current data. The models differ according to how they characterize the cognition mechanisms by which item scores were achieved. Indeed, some models may be more appropriate for one type of psychological process than another (Sung & Kang, 2006). So, the question is whether or not one of these models faithfully reflects the psychological reality that produced the data (Nering & Ostini, 2010). For this, the nature of the processes underlying the test item responses must be clarified (Sung & Kang, 2006).

For the items, we think in terms of multiple steps that have to be taken in the solution process. The difference between the models is how these steps are taken. The PCM is an adjacent-categories logit model and thus models the conditional probability of response \( x \), given response \( x - 1 \) (Rijmen et al., 2003). Thus, when a child is able to apply a non-analogical duplication strategy, what is the probability that an analogical partially correct strategy will be applied. The GRM is a cumulative logits model and models the cumulative probability of a response of category \( x \) or higher (Rijmen et al., 2003). So, what is the probability of a child, able of applying a non-analogical strategy, to take the next step and apply an analogical strategy. The CRM is a continuation-ratio logit model and models the conditional probability of response \( x \), given response \( x \) or \( x - 1 \) (Rijmen et al., 2003). So, when a child is able to apply a non-analogical duplication strategy, what is the probability that an analogical strategy will be applied.

The model selected should reflect how the children actually responded to the test (Reise & Revicki, 2014). Clarifying the nature of the process that underlies the item responses in order to find the most appropriate model is however very challenging to do. It is therefore very difficult to differentiate polytomous IRT models based on philosophical criteria (Nering & Ostini, 2010).
4.3.5.4 Practical concerns

Some last notes about model selection involve the practical concerns that might play a role in the model selection decision (Embretson & Reise, 2000; Reise & Revicki, 2014). The PCM and GRM are applicable in many statistical software programs (including multiple packages in R, which is a free software environment for statistical computing (R Development Core Team, 2013)). Because there are multiple software packages available for these kind of models, a lot of information is provided about the PCM and GRM. In contrast, software for estimating sequential models is not widely available (Van der Ark, 2001). Since relatively little attention is been given to continuation ratio models, there is less information about these kinds of models available (Hemker et al., 2000).

4.3.6 Conclusion methodological research question

The partial credit model, graded response model and continuation ratio model were fitted to the common pretest items. In addition, the group structure was taken into account. Based on the model fit and interpretation of the parameters of the fitted models, features of the data and some practical concerns, the most appropriate model for the analysis of the current data is the graded response model. The main argument for this decision has been model fit. With use of this model, the substantive research question will be answered which is described in the next Section 4.4.
4.4 Substantial research question: ‘Are there training- and age related changes regarding the strategies children use to solve figural analogical problems?’

The second part of this study aimed to answer the substantial research question. In order to answer this question, all items were included in analysis (including the responses of the posttest-items). We started with a null model (based on Embretson’s approach) where after the person predictors were included stepwise. Since each model is nested in the previous model, these models are statistically compared using their deviances. This is referred to as a likelihood ratio test (LR-test; Singer & Willet, 2003). The fit statistics of the models and results of the LR-tests are presented in Table 9.

4.4.1 Proportion of strategy use per training condition

Table 8 shows the proportion of strategy use per training condition over the two test occasions. These proportions are graphically presented in Figure 14.

Table 8
Proportion of Strategy Use of Children by Training Condition and Test occasion

<table>
<thead>
<tr>
<th>Training condition</th>
<th>N</th>
<th>Test occasion</th>
<th>Non-analogical other</th>
<th>Non-analogical duplication</th>
<th>Analogical partially correct</th>
<th>Analogical correct</th>
</tr>
</thead>
<tbody>
<tr>
<td>Control</td>
<td>95</td>
<td>Pretest</td>
<td>0.19</td>
<td>0.17</td>
<td>0.49</td>
<td>0.15</td>
</tr>
<tr>
<td>Control</td>
<td></td>
<td>Posttest</td>
<td>0.19</td>
<td>0.10</td>
<td>0.45</td>
<td>0.27</td>
</tr>
<tr>
<td>Practice</td>
<td>279</td>
<td>Pretest</td>
<td>0.16</td>
<td>0.22</td>
<td>0.33</td>
<td>0.29</td>
</tr>
<tr>
<td>Practice</td>
<td></td>
<td>Posttest</td>
<td>0.13</td>
<td>0.20</td>
<td>0.30</td>
<td>0.37</td>
</tr>
<tr>
<td>Outcome feedback</td>
<td>202</td>
<td>Pretest</td>
<td>0.14</td>
<td>0.22</td>
<td>0.31</td>
<td>0.33</td>
</tr>
<tr>
<td>Outcome feedback</td>
<td></td>
<td>Posttest</td>
<td>0.08</td>
<td>0.10</td>
<td>0.30</td>
<td>0.53</td>
</tr>
<tr>
<td>Graduated prompts</td>
<td>431</td>
<td>Pretest</td>
<td>0.19</td>
<td>0.26</td>
<td>0.31</td>
<td>0.25</td>
</tr>
<tr>
<td>Graduated prompts</td>
<td></td>
<td>Posttest</td>
<td>0.09</td>
<td>0.17</td>
<td>0.31</td>
<td>0.49</td>
</tr>
</tbody>
</table>

For comparison, the proportion of applied solution strategies are calculated over the items that were administered in each training condition. This resulted in 16 items ranging in difficulty from 2 to 7 transformations. It can be concluded from Figure 14 that for all training conditions, the proportion of analogical correct strategies increased from pretest to posttest. In addition, (slight) decreases are visible for all the other solution strategies.
Comparability of the training conditions

As explained earlier, we aggregated the data of the different six experiments and used it for analyses in this study. In each experiment, children were paired based on age, gender, classroom and cognitive ability estimates, after which they were randomly assigned to different training conditions. However, aggregating the six datasets leads to the fact that the groups of children receiving different training conditions might not be comparable with each other. Figure 15 indicates that the ages of the children differed between the training conditions. Levene’s test showed the variances of age to be significantly different in the four training conditions.
(F(3, 1003) = 15.87, \( p < .001 \)). Thus equal variances was not assumed and therefore the Kruskal-Wallis test was performed, which is an alternative for the one-way analysis of variance (ANOVA). Results from this test showed that the training conditions differed with regard to the average age of the children (\( H(3) = 148.76, p < .001 \)). Pairwise comparisons with adjusted \( p \)-values showed that there were significant differences in age between training conditions. In the control condition, age differed significantly compared to practice, outcome feedback and graduated prompts (\( p < .001, r = -.39; p < .001, r = -.45 \) and \( p < .001, r = -.71 \) respectively). There were also significant differences in age between graduated prompts and practice compared to outcome feedback (\( p < .001, r = .24 \) and \( p < .001, r = -.24 \) respectively). Finally, no significant differences were found between the graduated prompts and practice condition (\( p = 1.00, r = .01 \)).

![Figure 15. Boxplots of age in years per training condition.](image)

### 4.4.3 Embretson’s Approach

In order to incorporate the relationship between the latent trait levels at both test occasions (pretest and posttest), a multidimensional model is necessary. Embretson’s approach (1991) will therefore be used. Following Embretson’s MRMLC, the first dimension is assumed to be the latent trait level at time of the pretest (\( \theta_0 \)) and the second dimension is a modifiability (\( \theta_1 \)) which refers to an additional ability uniquely present at the posttest. Of course, this approach of the MRMLC will be generalized to the graded response model to accommodate the polytomously scored items in this study.

Since the items of the pre- and posttest are isomorphs, they had to be solved using the same transformations. The number of transformations is related to the item difficulty (Stevenson et al., 2013). We assume that the item difficulty of two isomorphs is equal and thus
constant over test occasions. This is referred to as measurement invariance since we want the change in the measurement of the strategies used to reflect the actual change in the underlying trait variable of a person after training (Millsap, 2010). If this change is due to the fact that the training conditions affect measurement qualities, measurement invariance does not hold (Adèr, Mellenbergh & Hand, 2008). Thus, constraints were incorporated on the item parameters by setting the between category threshold parameters of the pretest items equal to the parameters of the corresponding posttest items. Item 511 was excluded from analysis since this item was only administered once during the posttest and therefore did not have a corresponding pretest item. Also, since none of the participants had used an analogical correct strategy at item 602 and 710 during the pretest, these response categories had zero responses. This resulted in two between category threshold parameters instead of three and therefore they could not be set equal to the parameters of the corresponding posttest items. Excluding these items resulted in a total of 64 items available for analysis (32 pretest and 32 posttest items).

In addition, following Embretson’s approach (1991), the items received equal discrimination parameters on the first dimension and unique discrimination parameters on the second dimension (Von Davier, Xu & Carstensen, 2009). The two dimensions were allowed to be correlated with each other.

4.4.4 Latent regression

Since we are interested in the effects of training and age on the strategies in analogical reasoning, the model was extended with the explanatory person predictors training condition and age, resulting in a latent regression. In the latent regression model, the \( \theta \) effects are decomposed into fixed-effect components in order to explain the differences between individuals; \( \theta = X\beta + \epsilon \) whereby X is a design matrix containing person-level covariate data (Chalmers, 2012). So in the current analyses, the person predictors are regressed on one or both dimensions in order to explain the children’s initial ability level and performance change (De Boeck & Wilson, 2004). There were no missing values on the person predictors and therefore all 1007 children were included in analyses.

4.4.5 Null model

The first model fitted to the data was the graded response model with constraints on the item parameters (M0). Since the data is an aggregation of six different experiments, the effect of training condition is largely depended on the experiment it was applied in. In addition, not
every training condition was used in every experiment, leading to an unbalanced design. In order to control for this experiment effect, the variable was added to the null model as a fixed regression effect. This way, each experiment has its own mean initial trait level and modifiability. The variable was dummy-coded with experiment 20091 as reference category. The R-code is presented in Appendix 4. From this R-code it is clear to see that the first dimension (or latent factor) affects all items (both the pretest and the posttest items). The second dimension only affects the posttest items. Therefore, the responses on the posttest items depend on the composite ability level of the initial ability level plus the modifiability. The effect of experiment was added to the model on both dimensions, since it could influence both the initial ability as well as the modifiability. In addition, the latent factors were allowed to be correlated with each other.

It was assumed that the participants were sampled from a normal population distribution with mean $\theta$ equal to 0 and standard deviation equal to 1. Thus, the mean of the initial ability level was fixed to zero (for identification purposes). In addition, the mean of the modifiability (the performance change from pretest to posttest) was estimated freely.

4.4.6 Person predictors

4.4.6.1 Effect of training condition

The person predictor training condition was regressed on the second dimension since we want to examine whether the performance change from pretest to posttest might be influenced by the type of training a child had received. The R-code for this model (M1) is presented in Appendix 4.2. Training condition was dummy coded with the graduated prompts training as reference category. The LR-test in Table 9 shows that M1 is a significant improvement of M0. Results showed that all training conditions (control, practice and the outcome feedback training) led to significantly less improvement in analogical reasoning skills compared to the graduated prompts training. Thus, performance change from pretest to posttest was influenced by training condition.

However, the initial group comparison showed that the ages of the children receiving the different training conditions were not equal. Therefore, the effect of training condition might be contaminated by the fact that the children receiving a certain training condition were much older or younger compared to another training condition. Therefore, the person predictor age will be added to the model in the next subsection.
4.4.6.2 Effect of age

The person predictor age was regressed on both dimensions since age could have an effect on initial ability level as well as on the performance change from pretest to posttest. The R-code for this model (M2) is presented in Appendix 4.3. Table 9 shows that M2 is a significant improvement of M1. Age had a significant influence on both the initial ability and the modifiability and will therefore be included in further analysis.

4.4.6.3 Interaction effect between training condition and age

In addition to the main effects of training condition and age, an interaction effect between both predictors was added to M2 resulting in the third model (M3). The R-code for this model (M3) is presented in Appendix 4.4. This way it was investigated whether older children benefitted more from a certain training type than younger children. Adding this interaction effect improved model fit as can be seen in Table 9.

Table 9
Fit statistics and LR-test of the Estimated Models

<table>
<thead>
<tr>
<th>Model</th>
<th>Nested Model</th>
<th>Fixed predictor effects</th>
<th>Deviance</th>
<th>AIC</th>
<th>BIC</th>
<th>#p</th>
<th>df</th>
<th>(\chi^2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>M0</td>
<td>Experiment</td>
<td></td>
<td>76432.1</td>
<td>76776.1</td>
<td>77621.4</td>
<td>172</td>
<td></td>
<td></td>
</tr>
<tr>
<td>M1</td>
<td>M0</td>
<td>+ training condition</td>
<td>76381.4</td>
<td>76731.4</td>
<td>77591.5</td>
<td>175</td>
<td>3</td>
<td>50.7***</td>
</tr>
<tr>
<td>M2</td>
<td>M1</td>
<td>+ age</td>
<td>76099.8</td>
<td>76453.8</td>
<td>77323.7</td>
<td>177</td>
<td>2</td>
<td>281.6***</td>
</tr>
<tr>
<td>M3</td>
<td>M2</td>
<td>+ training condition*age</td>
<td>76077.4</td>
<td>76437.4</td>
<td>77322.1</td>
<td>180</td>
<td>3</td>
<td>22.4***</td>
</tr>
</tbody>
</table>

*p < .05, **p < .01, ***p < .001.
4.4.7 Final model and interpretation

The model that is selected as the final model is M3 with significant fixed effects for experiment, training condition, age and an interaction between training condition and age. The parameter estimates of this model are presented in Table 10.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Initial ability</th>
<th>Modifiability</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\beta$ (SE)</td>
<td>$p$-value</td>
</tr>
<tr>
<td>Experiment</td>
<td></td>
<td></td>
</tr>
<tr>
<td>20091 (Reference)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>20092</td>
<td>-0.19 (0.14)</td>
<td>.077</td>
</tr>
<tr>
<td>20101</td>
<td>-0.58 (0.14)</td>
<td>.000</td>
</tr>
<tr>
<td>20102</td>
<td>-1.31 (0.07)</td>
<td>.000</td>
</tr>
<tr>
<td>20111</td>
<td>-1.32 (0.07)</td>
<td>.000</td>
</tr>
<tr>
<td>20121</td>
<td>-1.08 (0.09)</td>
<td>.000</td>
</tr>
<tr>
<td>Training condition</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Control</td>
<td>-0.60 (0.11)</td>
<td>.000</td>
</tr>
<tr>
<td>Practice</td>
<td>-0.64 (0.08)</td>
<td>.000</td>
</tr>
<tr>
<td>Outcome feedback</td>
<td>-0.57 (0.31)</td>
<td>.035</td>
</tr>
<tr>
<td>Graduated prompts (Reference)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Age</td>
<td>0.69 (0.04)</td>
<td>.000</td>
</tr>
<tr>
<td>Training condition*age</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Control*age</td>
<td>0.32 (0.07)</td>
<td>.000</td>
</tr>
<tr>
<td>Practice*age</td>
<td>0.06 (0.06)</td>
<td>.156</td>
</tr>
<tr>
<td>Outcome feedback*age</td>
<td>0.59 (0.27)</td>
<td>.016</td>
</tr>
<tr>
<td>Graduated prompts*age (Reference)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The parameter estimates of the experiments showed differences in initial ability and modifiability, compared to the reference category 20091. Although these effects are partly due to the differences in age, they cannot be totally controlled for by age alone. Other factors such as intelligence could also play a role in the differences in initial ability and modifiability.

As can be seen from Table 10, the effect of training condition was regressed on the modifiability. Controlling for the effects of age and experiment, all training conditions (control, practice and outcome feedback) led to significant less improvement in solving analogical problems compared to the graduated prompts training ($\beta = -0.60$, $SE = 0.11$, $p < .001$; $\beta = -0.64$, $SE = 0.08$, $p < .001$ and $\beta = -0.57$, $SE = 0.31$, $p = .035$ respectively). Thus, children trained with
the graduated prompts technique showed greater performance change from pretest to posttest compared to the other training conditions.

The effects of experiment as well as training condition are both overall effects since these are based on groups of children. The individual effect of age showed to be a significant predictor of both initial analogical reasoning skills and performance change. The results showed older children to have higher initial ability scores than younger children ($\beta = 0.69, \text{SE} = 0.04, p < .001$). When a child increased in age with one unit, his or her analogical reasoning skills increased with 0.69. Between age and modifiability a negative relation was found ($\beta = -0.37, \text{SE} = 0.05, p < .001$) indicating that older children showed less improvement from pretest to posttest than younger children.

As previously mentioned, the graduated prompts training led to the greatest performance change compared to the other training conditions. The positive interaction effects between training condition and age led to the fact that the differences in modifiabilities between the training conditions were larger for younger children than for older children. Although the graduated prompts training led to the highest modifiability, older children performed significantly better after being in the control or outcome feedback condition compared to younger children ($\beta = 0.32, \text{SE} = 0.07, p < .001; \beta = 0.59, \text{SE} = 0.27, p = .016$ respectively). For children in the practice condition, this effect was not significant ($\beta = 0.06, \text{SE} = 0.06, p = .156$).

For illustration purposes, let us assume that for a particular child, the centered variable age would take on the value -1.0 and this child was in experiment 20101. When this child was trained in the control condition, his or her modifiability would equal 0.35 ($\text{modifiability} = 0 + 0.90*\text{experiment 20101} – 0.60*\text{control} – 0.37*\text{age} + 0.32*\text{age}\times\text{control}$). When the same child received the graduated prompts training, his or her modifiability would equal 1.27 ($\text{modifiability} = 0 + 0.90*\text{experiment 20101} – 0.37*\text{age}$). Similarly, the modifiability of a child with a centered age value of 1.0 in the control condition would equal 0.25 and the same child trained with graduated prompts would have a modifiability of 0.53. The difference in modifiability between both training conditions was larger for the younger child (age = -1.0, $\Delta_{\text{GP–Control}} = 0.92$) than for the older child (age = 1.0. $\Delta_{\text{GP–Control}} = 0.28$). This indicates that younger children benefitted more than older children from the more comprehensive training condition compared to the other three training conditions.

Finally, a small negative correlation was found between the children’s initial ability and modifiability ($r = -0.34$). This indicates that children with lower initial ability scores tended to show greater improvement than children with higher initial abilities scores.
5. Discussion

Responses of children on a figural analogy task were collected in a total of six experiments. Each experiment had a pretest-training-posttest design whereby children were randomly assigned to one or two out of four different training conditions. The present study had a methodological as well as a substantive aim. The methodological aim was to find the most appropriate model from three families of polytomous IRT models for the analyses of the current data. The substantive aim was to gain insight into the effect training and age might have on the strategies children used in order to solve the figural analogy problems.

5.1 Model selection

The first research question concerned the most appropriate polytomous IRT model for the analyses of the current data. Three models were taken into consideration which were the PCM, GRM and CRM. After fitting these models to the common pretest items, the GRM was selected as most appropriate model. Nowadays, there are many potentially applicable models available. Therefore it can be very challenging for a researcher to find the model that is most appropriate for his or her dataset and research intentions. This research provided insights into the choices and considerations present in model selection. Model selection was based on the fit indices, interpretation of the parameters and substantial features of the data. Since there were only little differences between the three models, the decision was mainly based on model fit. Since the GRM is an extension of the 2PL model, this model included an item’s difficulty parameter as well as an item discrimination parameter. Therefore, the number of parameters was always higher for the GRM compared to the other two models. Generally speaking, models with fewer parameters are better than models with more parameters because than the sample size needed for estimation will be smaller (Embretson & Reise, 2000). However, the items showed to have differing discriminations and therefore the GRM was more appropriate then models without this discrimination parameter. In addition, this study had a rather large sample size.

However, in search of the most appropriate model, it is implicitly assumed that the correct model for this type of data is among the models considered (Akkermans, 1998; Sung & Kang, 2006; Agresti, 2013). When only two out of these three models were considered, one of these two models would have fitted best. This would indicate that that particular model is the most appropriate one, while in fact the third uninvestigated model might be the most appropriate. Thus, one should keep in mind that the graded response model appeared to be the
most appropriate in this study, but this does not conclude that this model is indeed the best model. On the other hand, Akkermans (1998) investigated the possibility to distinguish between data that was generated under different models. In her dissertation, she examined the differences between item response models from different families and the possibility to distinguish between them. The families of models of her focus were the family of partial credit models, the family of graded response models and the family of sequential model. In order to decide between the three models, a zero/one loss function and a maximum likelihood decision strategy were used. So, the model with the largest likelihood would be selected. Akkermans (1998) found that the PCM and the SM, both fitted to data from the GRM, appeared to have on average approximately equal likelihoods. In addition, as previously mentioned, Nering and Ostini (2010) stated that there has been little evidence for polytomous IRT models to produce substantially different measurement outcomes when applied to the same data. This indicates that the practical consequences of choosing a ‘wrong’ model might not be very large (Akkermans, 1998). Our results confirm this finding.

5.2 Effect of training and age on the change of strategy use

Analyses on the second research question provided insight into the effect of training and age on children’s progression of strategy change. Both training and age were found to be important predictors of children’s analogical reasoning skills. When looking at the proportion of strategy use per test occasion, the use of analogical correct strategies increased from pretest to posttest. After fitting a multidimensional graded response model to the data, results showed that, controlling for age and experiment, the graduated prompts training condition led to significantly greater performance change from pretest to posttest compared to the control, practice and outcome feedback training.

Age turned out to be an important factor in the prediction of analogical reasoning skills. Older children were found to have higher initial ability scores than younger children. This result is in line with previous research, which showed that older children are better at solving analogical problems than younger children (e.g., Hosenfeld, van den Boom & Resing, 1997). A negative relation was found between age and modifiability, meaning that younger children improved more after training than older children. Although a similar experiment from Stevenson et al. (2013) found no relation between the children’s degree of improvement from pretest to posttest and their age, they did find improvement after training to be influenced by initial ability. In their research, children with lower initial ability scores showed greater
performance change from pretest to posttest. Thus the results found in our study, that younger children had lower initial ability scores and showed more improvement after training, is in line with our expectations.

The interaction effect between training condition and age indicated that younger children benefitted more than older children from the graduated prompts training compared to the control and outcome feedback training. The fact that younger children compared to older children benefit more from a more comprehensive training is very useful and relevant information for teachers or educational psychologists. Further research on this topic can give more precise information on which training/teaching style can be best applied and at which age, so that children can maximize their analogical reasoning skills with the most efficient technique.

5.3 Limitations

Previously, it was discussed that children from different experiments were assumed to come from different populations with potentially different ability distributions. In the final model, the experiments were included as fixed regression effects. Hereby, each experiment had its own mean initial trait level and modifiability but it was assumed that the variances of the latent ability distributions were equal. A reason why the experiments were not included in our final model as random effects was because the use of latent regression random effects was currently disabled in the mirt package (Chalmers, 2012). Table 6 showed the within-experiment variances for the graded response model to differ slightly from each other. In order to examine whether the within-experiment variances differ significantly from each other, the same model was fitted to the data but the variances were constrained to be equal. This model was a significant improvement of the model with freely estimated variances ($\chi^2 (5, n = 1007) = 39.51, p < .001$). Therefore, it would have been ideal if we could have included the experiments as random effects in the latent regression model.

5.4 Methodological considerations

The fact that the current data is an aggregation of six experiments has led to a quite complex dataset. Six samples of children with varying ages were randomly assigned to one or two out of the four training conditions and were administered a set of items. The set of items administered differed with each experiment, whereby seven items were administered in every experiment. Thus, many item responses were missing by design. However, the use of the R-package mirt provided us with flexible software that is very capable of handling these missing
values. \textit{mirt} uses full information maximum likelihood estimation (FIML; Chalmers, 2012). With FIML estimation, a likelihood function is estimated for each individual based on all the information contained in the response pattern. Thus, all available information is used to estimate the model parameters (Forero & Maydeu-Olivares, 2009). So, even though not all items were administered in each experiment, model fit information is derived from a summation across fit functions for individual cases, and thus model fit information is based on all cases (Newsom, 2015). In addition, when fitting the final model on the common items so there was no missing data, similar results were found compared to the analyses on all items.

Embretson’s MRMLC was generalized to the graded response model to accommodate the polytomously scored items. In addition, the model was extended with two explanatory person predictors. Because the psychometric properties of the test scores should not change from the pretest to the posttest, we put equality constraints on the item parameters in order to hold the assumption of measurement invariance (Millsap, 2010). Therefore it is important to keep in mind that the conclusions that were drawn from this research, were made under the assumption of measurement invariance.

In order to determine whether the item response functions for two corresponding items across test occasion (pretest and posttest) were the same, a model without equality constraints was fitted to the data. This model improved model fit ($\chi^2 (128, n = 1007) = 1181.03, p < .001$), indicating there were items changing over time in their item parameters (Millsap, 2010). Of course, ideally, we would want the items to measure the same at both occasions. However, one could not have expected the item parameters to be exactly the same, since the posttest is a two-dimensional construct with the modifiability playing an important role. Results from the model without equality constraints does not lead to different conclusions.

### 5.5 Recommendations for future research

In this study, model selection was based upon the fit indices, interpretation of the parameters and substantial features of the data. However, model selection could also have been based on different or additional methods (Akkermans, 1998). Akkermans (1998) reported a different way of choosing between models from different families. She calculated the ideal observer index (IOI) and showed its strength in differentiating between the GPCM and GRM (Sung & Kang, 2006). However, as mentioned in Sung and Kang (2006), the IOI is according to Ostini and Nering (2005) not a practical method for selecting the most appropriate model since it can only be estimated with simulation data.
Sung and Kang (2006) do report some additional model selection methods that could be used in future research. They mention methods like the DIC, BF and/or CVLL that could be taken into account when deciding among models (Sung & Kang, 2006).

As previously mentioned, the current conclusions are made based on the assumption that the variances of the latent ability distributions were equal. Since the use of latent regression random effects was currently disabled in mirt, there was no other option within this package than to include the effects of the experiments as fixed effects. However, it would be very interesting to see whether the results would have led to different conclusions when this option was possible.

Unfortunately, we were not able to test whether the same construct was being measured across the six experiments in the final model. Since the experiments all had a rather different sample of children, it would be interesting to investigate in future research whether measurement invariance holds for the six experiments and thus whether the same construct was measured across the experiments.

5.6 Conclusions

In the present study, two research questions were answered. The first and methodological research question concerned the selection of the most appropriate model from three polytomous IRT models (PCM, GRM and CRM) for the analysis of the current data. This research provided insight into the choices and considerations one could have in model selection. All three models were fitted to the common pretest items of a figural analogy task. Based on the fit indices (AIC, BIC and deviance), interpretation of the parameters and substantial features of the data, the GRM was selected as the most appropriate model for the analysis of the current data. It would be interesting to find out if this same model would have been selected when additional models where applied and alternative methods for model selection were used.

The second and substantive research question concerned the effects of training condition and age on children’s use of solution strategies at the pre- and posttest. Age was found to be an important predictor in children’s initial ability. In addition, age as well as training condition and their interaction were all significant predictors of children’s performance change from pretest to posttest. These results all provide insight into the development of cognitive capacity and potential of children to apply knowledge, obtained in one context, in a new situation. Since the focus of this study was on the solution strategies children used in order to solve the figural analogy problems, many insight have been provided into the strategic development. Especially
the interaction between age and training condition is very relevant for educational psychologists and teachers. It would therefore be very useful if future research focuses on this interaction to find out more about the type of training or teaching that is most beneficial for children to reach their full potential in a certain age-group.
6. Literature


Appendix

1. R-codes of PCM, GRM and CRM fitted on the common pretest items

1.1 PCM

1.1.1 R-code PCM

```r
> head(DATA_pretest)
   i1201 i1204 i1301 i1404 i1502 i1505 i1604
1     3     3     2     2     2     0     2
2     3     3     2     2     2     2     2
3     3     3     3     3     3     3     2
4     3     3     3     3     3     2     3
5     3     3     3     3     2     2     2
6     3     3     3     3     2     2     2

PCM <- mirt(data=DATA_pretest, 1, itemtype='Rasch', SE=TRUE)
> PCM
Full-information item factor analysis with 1 factor(s).
Converged within 1e-04 tolerance after 58 EM iterations.
mirt version: 1.8
M-step optimizer: BFGS
EM acceleration: Ramsay
Number of rectangular quadrature: 41

Information matrix estimated with method: crossprod
Condition number of information matrix = 126.3544
Second-order test: model is a possible local maximum

Log-likelihood = -7854.339
AIC = 15752.68; AICc = 15753.71
BIC = 15860.8; SABIC = 15790.93

> coef(PCM, IRTpars=TRUE)
$\text{i1201}$
   a b1 b2 b3
par 1 2.737 0.684 0.186

$\text{i1204}$
   a b1 b2 b3
par 1 3.413 0.074 0.354

$\text{i1301}$
   a b1 b2 b3
par 1 2.584 0.824 0.4

$\text{i1404}$
   a b1 b2 b3
par 1 1.002 0.686 1.481

$\text{i1502}$
   a b1 b2 b3
par 1 -0.135 -0.135 2.145

$\text{i1505}$
   a b1 b2 b3
par 1 -0.246 -0.004 2.089

$\text{i1604}$
   a b1 b2 b3
par 1 0.099 -0.117 2.419
$\text{GroupPars}$
   MEAN 1 COV 11
par 0 1.527
1.1.2 Estimated slope intercept parameters of the PCM

Table A1

<table>
<thead>
<tr>
<th>Item</th>
<th>$d_1$ (SE)</th>
<th>$d_2$ (SE)</th>
<th>$d_3$ (SE)</th>
</tr>
</thead>
<tbody>
<tr>
<td>201</td>
<td>2.74 (0.19)</td>
<td>3.42 (0.20)</td>
<td>3.24 (0.23)</td>
</tr>
<tr>
<td>204</td>
<td>3.41 (0.21)</td>
<td>3.34 (0.23)</td>
<td>3.69 (0.25)</td>
</tr>
<tr>
<td>301</td>
<td>2.58 (0.17)</td>
<td>3.41 (0.19)</td>
<td>3.01 (0.22)</td>
</tr>
<tr>
<td>404</td>
<td>1.00 (0.13)</td>
<td>1.69 (0.14)</td>
<td>0.21 (0.19)</td>
</tr>
<tr>
<td>502</td>
<td>0.14 (0.11)</td>
<td>0.27 (0.14)</td>
<td>-1.87 (0.20)</td>
</tr>
<tr>
<td>505</td>
<td>0.25 (0.12)</td>
<td>0.25 (0.14)</td>
<td>-1.84 (0.20)</td>
</tr>
<tr>
<td>604</td>
<td>-0.10 (0.12)</td>
<td>0.02 (0.13)</td>
<td>-2.40 (0.20)</td>
</tr>
</tbody>
</table>

1.2 GRM

1.2.1 R-code GRM

```r
GRM <- mirt(data=DATA_pretest, 1, itemtype='graded', SE=TRUE)

> GRM
```

Full-information item factor analysis with 1 factor(s).
Converged within 1e-04 tolerance after 25 EM iterations.
mirt version: 1.8
M-step optimizer: BFGS
EM acceleration: Ramsay
Number of rectangular quadrature: 41

Information matrix estimated with method: crossprod
Condition number of information matrix = 111.0331
Second-order test: model is a possible local maximum

Log-likelihood = -7766.516  
AIC = 15589.03; AICc = 15590.69  
BIC = 15726.64; SABIC = 15637.71

```r
> coef(GRM, IRTpars=TRUE)
```

<table>
<thead>
<tr>
<th>Item</th>
<th>a</th>
<th>b1</th>
<th>b2</th>
<th>b3</th>
</tr>
</thead>
<tbody>
<tr>
<td>1201</td>
<td>1.779</td>
<td>-2.19</td>
<td>-0.703</td>
<td>0.399</td>
</tr>
<tr>
<td>1204</td>
<td>1.506</td>
<td>-2.732</td>
<td>-0.45</td>
<td>0.314</td>
</tr>
<tr>
<td>1301</td>
<td>1.019</td>
<td>-3.097</td>
<td>-0.981</td>
<td>0.739</td>
</tr>
<tr>
<td>1404</td>
<td>2.049</td>
<td>-1.111</td>
<td>-0.29</td>
<td>1.195</td>
</tr>
<tr>
<td>1502</td>
<td>3.125</td>
<td>-0.507</td>
<td>0.135</td>
<td>1.482</td>
</tr>
<tr>
<td>1505</td>
<td>2.24</td>
<td>-0.562</td>
<td>0.21</td>
<td>1.601</td>
</tr>
<tr>
<td>1604</td>
<td>1.938</td>
<td>-0.452</td>
<td>0.257</td>
<td>1.863</td>
</tr>
</tbody>
</table>
1.2.2 Estimated slope intercept parameters of the GRM

Table A2

<table>
<thead>
<tr>
<th>Item</th>
<th>(a) (SE)</th>
<th>(d_1) (SE)</th>
<th>(d_2) (SE)</th>
<th>(d_3) (SE)</th>
</tr>
</thead>
<tbody>
<tr>
<td>201</td>
<td>1.78 (0.12)</td>
<td>3.89 (0.22)</td>
<td>1.25 (0.11)</td>
<td>-0.71 (0.11)</td>
</tr>
<tr>
<td>204</td>
<td>1.51 (0.11)</td>
<td>4.11 (0.23)</td>
<td>0.68 (0.10)</td>
<td>-0.47 (0.10)</td>
</tr>
<tr>
<td>301</td>
<td>1.02 (0.08)</td>
<td>3.15 (0.16)</td>
<td>1.00 (0.09)</td>
<td>-0.75 (0.09)</td>
</tr>
<tr>
<td>404</td>
<td>2.05 (0.14)</td>
<td>2.28 (0.14)</td>
<td>0.60 (0.13)</td>
<td>-2.45 (0.15)</td>
</tr>
<tr>
<td>502</td>
<td>3.13 (0.23)</td>
<td>1.59 (0.17)</td>
<td>-0.42 (0.16)</td>
<td>-4.63 (0.32)</td>
</tr>
<tr>
<td>505</td>
<td>2.24 (0.15)</td>
<td>1.26 (0.13)</td>
<td>-0.47 (0.12)</td>
<td>-3.59 (0.21)</td>
</tr>
<tr>
<td>604</td>
<td>1.94 (0.13)</td>
<td>0.88 (0.11)</td>
<td>-0.50 (0.12)</td>
<td>-3.61 (0.19)</td>
</tr>
</tbody>
</table>

1.3 CRM

1.3.1 R-code CRM

```r
linmap <- cbind(c(0,1,1,1),c(NA,0,1,1),c(NA,NA,0,1))
> linmap
  [,1] [,2] [,3]
[1,]  0   NA   NA
[2,]  1    0   NA
[3,]  1    1    0
[4,]  1    1    1

DATA_pretest.mat <- as.matrix(DATA_pretest)
DATA_dendrify <- dendrify(DATA_pretest.mat,linmap)
# Dendrify expands a wide-form matrix of item responses to a long-form data frame of sub-item responses
# Including experiment and corresponding dummy variables:
> head(DATA_dendrify)
  value item person node sub experiment e20092 e20101 e20102 e20111 e20121
1   1     1   p0001 node1 1:node1 20121    0    0    0    0    1
2   1     1   p0002 node1 1:node1 20121    0    0    0    0    1
3   1     1   p0003 node1 1:node1 20121    0    0    0    0    1
4   1     1   p0004 node1 1:node1 20121    0    0    0    0    1
5   1     1   p0005 node1 1:node1 20121    0    0    0    0    1
6   1     1   p0006 node1 1:node1 20121    0    0    0    0    1

CRM <- glmer(value ~ 0 + item:node + (1|person), family = binomial,
  data=DATA_dendrify, control= glmerControl(optimizer="bobyqa"))
> summary(CRM)
```

Generalized linear mixed model fit by maximum likelihood (Laplace Approximation) ['glmerMod']
  Family: binomial ( logit )
  Formula: value ~ 0 + item:node + (1 | person)
  Data: DATA_dendrify
  Control: glmerControl(optimizer = "bobyqa")

```
AIC  BIC logLik deviance df.resid
15668.4 15837.9  -7812.2 15624.4  16375
```
Scaled residuals:

<table>
<thead>
<tr>
<th></th>
<th>Min</th>
<th>1Q</th>
<th>Median</th>
<th>3Q</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>-19.6190</td>
<td>-0.5496</td>
<td>0.2016</td>
<td>0.4979</td>
<td>6.2571</td>
</tr>
</tbody>
</table>

Random effects:

<table>
<thead>
<tr>
<th>Groups</th>
<th>Name</th>
<th>Variance</th>
<th>Std.Dev.</th>
</tr>
</thead>
<tbody>
<tr>
<td>person</td>
<td>(Intercept)</td>
<td>2.543</td>
<td>1.595</td>
</tr>
</tbody>
</table>

Number of obs: 16397, groups: person, 1002

Fixed effects:

| Estimate | Std. Error | z value | Pr(>|z|) |
|----------|------------|---------|---------|
| item1:nodenode1 | 3.78591 | 0.16342 | 23.167 < 2e-16 *** |
| item2:nodenode1 | 4.24876 | 0.19201 | 22.128 < 2e-16 *** |
| item3:nodenode1 | 3.67553 | 0.15762 | 23.318 < 2e-16 *** |
| item4:nodenode1 | 2.02277 | 0.10539 | 19.193 < 2e-16 *** |
| item5:nodenode1 | 0.92915 | 0.09350 | 9.937 < 2e-16 *** |
| item6:nodenode1 | 0.98417 | 0.09384 | 10.488 < 2e-16 *** |
| item7:nodenode1 | 0.68933 | 0.09238 | 7.462 8.54e-14 *** |
| item1:nodenode2 | 1.29179 | 0.09825 | 13.147 < 2e-16 *** |
| item2:nodenode2 | 0.77497 | 0.09379 | 8.263 < 2e-16 *** |
| item3:nodenode2 | 1.39066 | 0.09965 | 13.956 < 2e-16 *** |
| item4:nodenode2 | 0.99537 | 0.10373 | 9.596 < 2e-16 *** |
| item5:nodenode2 | 0.30205 | 0.11007 | 2.744 0.006065 ** |
| item6:nodenode2 | 0.22205 | 0.10880 | 2.041 0.041251 * |
| item7:nodenode2 | 0.39412 | 0.11401 | 3.457 0.000546 *** |
| item1:nodenode3 | -0.23298 | 0.10557 | -2.207 0.027321 * |
| item2:nodenode3 | 0.37317 | 0.11352 | 3.287 0.001012 ** |
| item3:nodenode3 | -0.39369 | 0.10542 | -3.734 0.000188 *** |
| item4:nodenode3 | -1.83040 | 0.11996 | -15.258 < 2e-16 *** |
| item5:nodenode3 | -2.65529 | 0.14333 | -18.526 < 2e-16 *** |
| item6:nodenode3 | -2.56692 | 0.14432 | -17.787 < 2e-16 *** |
| item7:nodenode3 | -2.90880 | 0.15339 | -18.963 < 2e-16 *** |

---

Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

2. R-codes of PCM2, GRM2 and CRM2 fitted on the common pretest items with multiple groups

2.1 R-code PCM2

```r
# for multigroup analysis, item responses and person covariates need to be ordered by group
group <- sort(group, index.return=T)
sort_results <- sort(group, index.return=T)
sort_results$six <- sort_results$x

DATA_pretest <- DATA_pretest[sort_results$six,]

PCM2 <- multipleGroup(data=DATA_pretest, l, group=group, invariance=c('slopes', 'intercepts', 'free_varcov', 'free_means'), itemtype='Rasch', SE=TRUE)
```

> PCM2

Full-information item factor analysis with 1 factor(s). Converged within 1e-04 tolerance after 137 EM iterations.
mirt version: 1.8
M-step optimizer: BFGS
EM acceleration: Ramsay
Number of rectangular quadrature: 41
Information matrix estimated with method: crossprod
Condition number of information matrix = 78.90254
Second-order test: model is a possible local maximum

Log-likelihood = -7564.064
AIC = 15192.13; AICc = 15194.3
BIC = 15349.4; SABIC = 15247.77

> coef(PCM2, IRTpars=TRUE)$`20091`
$par 1 $i1201
  a b1 b2 b3
par 1 -1.888 0.044 0.911
$i1204
  a b1 b2 b3
par 1 -2.56 0.796 0.365
$i1301
  a b1 b2 b3
par 1 -1.739 -0.097 1.129
$i1404
  a b1 b2 b3
par 1 -0.227 0.019 2.251
$i1502
  a b1 b2 b3
par 1 0.601 0.567 2.964
$i1505
  a b1 b2 b3
par 1 0.492 0.699 2.909
$i1604
  a b1 b2 b3
par 1 0.83 0.588 3.253
$GroupPars
  MEAN_1 COV_11
par 0 0.344

> coef(PCM2, IRTpars=TRUE)$`20092`$GroupPars
  MEAN_1 COV_11
par 0.371 0.231
> coef(PCM2, IRTpars=TRUE)$`20101`$GroupPars
  MEAN_1 COV_11
par 1.187 0.249
> coef(PCM2, IRTpars=TRUE)$`20102`$GroupPars
  MEAN_1 COV_11
par 0.277 0.739
> coef(PCM2, IRTpars=TRUE)$`20111`$GroupPars
  MEAN_1 COV_11
par 0.327 1.083
> coef(PCM2, IRTpars=TRUE)$`20121`$GroupPars
  MEAN_1 COV_11
par 2.317 1.072

2.2 R-code GRM2

GRM2 <- multipleGroup(data=DATA_pretest, 1, group=group,
invariance=c('slopes', 'intercepts', 'free_varcov', 'free_means'),
itemtype='graded', SE=TRUE)

> GRM2
Full-information item factor analysis with 1 factor(s).
Converged within 1e-04 tolerance after 157 EM iterations.
mirt version: 1.8
STATISTICAL MODELS OF CHILDREN’S STRATEGY CHANGE IN ANALOGICAL REASONING

M-step optimizer: BFGS
EM acceleration: Ramsay
Number of rectangular quadrature: 41

Information matrix estimated with method: crossprod
Condition number of information matrix = 159.1784
Second-order test: model is a possible local maximum

Log-likelihood = -7484.973
AIC = 15045.95; AICc = 15049.01
BIC = 15232.71; SABIC = 15112.02

> coef(GRM2, IRTpars=TRUE)$'20091'
  $i1201
  a b1 b2 b3
  par 1.215 -2.303 -0.141 1.51
  $i1204
  a b1 b2 b3
  par 1.041 -3.054 0.242 1.38
  $i1301
  a b1 b2 b3
  par 0.66 -3.825 -0.586 2.062
  $i1404
  a b1 b2 b3
  par 1.328 -0.791 0.461 2.779
  $i1502
  a b1 b2 b3
  par 1.957 0.113 1.114 3.251
  $i1505
  a b1 b2 b3
  par 1.411 0.036 1.237 3.429
  $i1604
  a b1 b2 b3
  par 1.253 0.21 1.313 3.813
$GroupPars
  MEAN_1 COV_11
  par 0 1
  > coef(GRM2, IRTpars=TRUE)$'20092'$'GroupPars'
  MEAN_1 COV_11
  par 0.398 0.574
  > coef(GRM2, IRTpars=TRUE)$'20101'$'GroupPars'
  MEAN_1 COV_11
  par 1.643 0.378
  > coef(GRM2, IRTpars=TRUE)$'20102'$'GroupPars'
  MEAN_1 COV_11
  par 0.339 1.297
  > coef(GRM2, IRTpars=TRUE)$'20111'$'GroupPars'
  MEAN_1 COV_11
  par 0.456 1.812
  > coef(GRM2, IRTpars=TRUE)$'20121'$'GroupPars'
  MEAN_1 COV_11
  par 2.736 1.128
### 2.3 R-code CRM2

CRM2 <- glmer(value ~ 0 + item:node + e20092 + e20101 + e20102 + e20111 + e20121 + (1|e20092) + (1|e20101) + (1|e20102) + (1|e20111) + (1|e20121) + (1|person), family = binomial, data=DATA_dendrify, control=glmerControl(optimizer="bobyqa"))

> summary(CRM2)

Generalized linear mixed model fit by maximum likelihood (Laplace Approximation) ['glmerMod']
Family: binomial  ( logit )
Formula: value ~ 0 + item:node + e20092 + e20101 + e20102 + e20111 + e20121 + (1 | e20092) + (1 | e20101) + (1 | e20102) + (1 | 20111) + (1 | e20121) + (1 | person)
Data: DATA_dendrify
Control: glmerControl(optimizer = "bobyqa")

AIC BIC logLik deviance df.resid
15165.0 15411.5  -7550.5 15101.0  16365

Scaled residuals:
          Min       1Q   Median       3Q      Max
-15.4208  -0.5544   0.2002   0.4948   6.3526

Random effects:
Groups     Name     Variance  Std.Dev. 
person (Intercept)  1.300e+00 1.140e+00 
e20121 (Intercept)  1.242e-14 1.114e-07 
e20111 (Intercept)  0.000e+00 0.000e+00 
e20102 (Intercept)  2.708e-16 1.646e-08 
e20101 (Intercept)  1.652e-14 1.285e-08 
e20092 (Intercept)  4.137e-15 6.432e-08 
Number of obs: 16397, groups:  person, 1002; e20121, 2; e20111, 2; e20102, 2; e20101, 2; e20092, 2

Fixed effects:  Estimate Std. Error z value Pr(>|z|)
  e20092        0.516179   0.191183   2.700  0.006935 **
  e20101        1.574934   0.215390   7.312  2.63e-13 ***
  e20102        0.418084   0.136518   3.062  0.002195 **
  e20111        0.478973   0.137202   3.491  0.000481 ***
  e20121        2.983258   0.148190  20.131  < 2e-16 ***
  itemi1:nodenode1  2.718869   0.183462  14.820  < 2e-16 ***
  itemi2:nodenode1  3.163596   0.207553  15.242  < 2e-16 ***
  itemi3:nodenode1  2.612452   0.178734  14.616  < 2e-16 ***
  itemi4:nodenode1  1.005823   0.135613   7.251  4.15e-13 ***
  itemi5:nodenode1 -0.070081   0.134553 -0.527  0.594418
  itemi6:nodenode1 -0.156300   0.150471 -1.040  0.296469
  itemi7:nodenode1  0.620967   0.149429   4.160  4.30e-05 ***
  itemi1:nodenode2  0.297204   0.135613   2.209  0.027187 *
  itemi2:nodenode2 -0.219380   0.132128 -1.660  0.096842 .
  itemi3:nodenode2  0.392143   0.135613   2.892  0.003832 **
  itemi4:nodenode2  0.005157   0.139340   0.037  0.970474
  itemi5:nodenode2 -0.697307   0.146846  -4.749  2.05e-06 ***
  itemi6:nodenode2 -0.733444   0.145465  -5.027  4.85e-07 ***
  itemi7:nodenode2 -0.613389   0.149429  -4.105  4.05e-05 ***
  itemi1:nodenode3 -1.232264   0.143831  -8.567  < 2e-16 ***
  itemi2:nodenode3 -0.626823   0.148908  -4.209  2.56e-05 ***
  itemi3:nodenode3  1.385278   0.143741   9.637  < 2e-16 ***
  itemi4:nodenode3 -2.869579   0.157981 -18.164  < 2e-16 ***
### 3. Item plots

#### 3.1 Item plots of PCM

```r
> itemplot(PCM, 1)
```

![Trace lines for item 1](image1)

```r
> itemplot(PCM, 2)
```

![Trace lines for item 2](image2)
> itemplot(PCM, 3)

> itemplot(PCM, 4)

> itemplot(PCM, 5)
3.2 Item plots of GRM

> itemplot(GRM, 1)
> itemplot(GRM, 2)

> itemplot(GRM, 3)

> itemplot(GRM, 4)
> itemplot(GRM, 5)

Trace lines for item 5

> itemplot(GRM, 6)

Trace lines for item 6

> itemplot(GRM, 7)

Trace lines for item 7
3.3 Item plots of CRM

Item plot of item 201

Item plot of item 204

Item plot of item 301
Item plot of item 404

Item plot of item 502

Item plot of item 505
4. R-codes of the estimated models fitted on all items

4.1 Null model (M0)

```r
### Mirt.model with constraints, correlation and free mean F2
S0 <- 'F1 = 1-64
F2 = 33-64
COV = F1*F2
MEAN = F2
CONSTRAIN =
(1,33,d1),(2,34,d1),(3,35,d1),(4,36,d1),(5,37,d1),(6,38,d1),
(7,39,d1),(8,40,d1),(9,41,d1),(10,42,d1),(11,43,d1),(12,44,d1),
(13,45,d1),(14,46,d1),(15,47,d1),(16,48,d1),(17,49,d1),(18,50,d1),
(19,51,d1),(20,52,d1),(21,53,d1),(22,54,d1),(23,55,d1),(24,56,d1),
(25,57,d1),(26,58,d1),(27,59,d1),(28,60,d1),(29,61,d1),(30,62,d1),
(31,63,d1),(32,64,d1),
(1,33,d2),(2,34,d2),(3,35,d2),(4,36,d2),(5,37,d2),(6,38,d2),
(7,39,d2),(8,40,d2),(9,41,d2),(10,42,d2),(11,43,d2),(12,44,d2),
(13,45,d2),(14,46,d2),(15,47,d2),(16,48,d2),(17,49,d2),(18,50,d2),
(19,51,d2),(20,52,d2),(21,53,d2),(22,54,d2),(23,55,d2),(24,56,d2),
(25,57,d2),(26,58,d2),(27,59,d2),(28,60,d2),(29,61,d2),(30,62,d2),
(31,63,d2),(32,64,d2),
(1,33,d3),(2,34,d3),(3,35,d3),(4,36,d3),(5,37,d3),(6,38,d3),
(7,39,d3),(8,40,d3),(9,41,d3),(10,42,d3),(11,43,d3),(12,44,d3),
(13,45,d3),(14,46,d3),(15,47,d3),(16,48,d3),(17,49,d3),(18,50,d3),
(19,51,d3),(20,52,d3),(21,53,d3),(22,54,d3),(23,55,d3),(24,56,d3),
(25,57,d3),(26,58,d3),(27,59,d3),(28,60,d3),(29,61,d3),(30,62,d3),
(31,63,d3),(32,64,d3),
(1,33,a1),(2,34,a1),(3,35,a1),(4,36,a1),(5,37,a1),(6,38,a1),
(7,39,a1),(8,40,a1),(9,41,a1),(10,42,a1),(11,43,a1),(12,44,a1),
(13,45,a1),(14,46,a1),(15,47,a1),(16,48,a1),(17,49,a1),(18,50,a1),
(19,51,a1),(20,52,a1),(21,53,a1),(22,54,a1),(23,55,a1),(24,56,a1),
(25,57,a1),(26,58,a1),(27,59,a1),(28,60,a1),(29,61,a1),(30,62,a1),
(31,63,a1),(32,64,a1)'
```
> head(covdata)

<table>
<thead>
<tr>
<th></th>
<th>e20092</th>
<th>e20101</th>
<th>e20102</th>
<th>e20111</th>
<th>e20121</th>
<th>Control</th>
<th>Practice</th>
<th>Feedback</th>
<th>age</th>
<th>age_cent</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>8.68</td>
<td>1.41</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>8.71</td>
<td>1.44</td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>8.53</td>
<td>1.26</td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>8.91</td>
<td>1.64</td>
</tr>
<tr>
<td>5</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>8.84</td>
<td>1.57</td>
</tr>
<tr>
<td>6</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>9.65</td>
<td>2.37</td>
</tr>
</tbody>
</table>

M0 <- mixedmirt(data=DATA, covdata, mirt.model(S0), itemtype='graded',
technical = list(NCYCLES = 1e5), lr.fixed = list(F1 = ~ e20092 + e20101 + e20102 + e20111 + e20121, F2 = ~ e20092 + e20101 + e20102 + e20111 + e20121))

> M0

Full-information item factor analysis with 2 factor(s).
Converged within 0.001 tolerance after 654 MHRM iterations.
mirt version: 1.8
M-step optimizer: NR

Information matrix estimated with method: MHRM
Condition number of information matrix = 1460.839
Second-order test: model is a possible local maximum

Log-likelihood = -38216.05, SE = 0.206
AIC = 76776.1; AICC = 76847.46
BIC = 77621.44; SABIC = 77075.15
> coef(M0, printSE=T)

$i1201
   a1  a2  d1  d2  d3
par 0.966 0 3.982 1.714 -0.014
SE  0.053 NA 0.137 0.087 0.079

$i1202
   a1  a2  d1  d2  d3
par 0.407 0 3.917 2.129 1.145
SE  0.048 NA 0.157 0.091 0.077

$i1203
   a1  a2  d1  d2  d3
par 1.042 0 3.437 1.170 -0.465
SE  0.055 NA 0.128 0.085 0.079

$i1204
   a1  a2  d1  d2  d3
par 1.287 0 5.127 1.573 0.525
SE  0.065 NA 0.187 0.128 0.085

$i1205
   a1  a2  d1  d2  d3
par 0.673 0 3.538 1.415 -0.265
SE  0.045 NA 0.123 0.074 0.068

$i1206
   a1  a2  d1  d2  d3
par 1.155 0 3.566 1.820 -0.053
SE  0.074 NA 0.178 0.125 0.107

$i1207
   a1  a2  d1  d2  d3
par 1.753 0 5.000 2.913 1.490
SE  0.136 NA 0.275 0.220 0.193

$i1208
   a1  a2  d1  d2  d3
par 1.113 0 3.872 2.107 0.152
SE  0.063 NA 0.160 0.112 0.095

$i1209
   a1  a2  d1  d2  d3
par 1.597 0 5.137 2.004 -0.024
SE  0.081 NA 0.197 0.114 0.098

$i1210
   a1  a2  d1  d2  d3
par 1.27 0 3.935 1.999 0.560
SE  0.10 NA 0.192 0.151 0.139

$i1211
   a1  a2  d1  d2  d3
par 0.924 0 2.395 0.672 -2.205
SE  0.056 NA 0.109 0.084 0.108

$i1212
   a1  a2  d1  d2  d3
par 1.685 0 3.794 1.793 -0.648
SE  0.076 NA 0.149 0.117 0.099

$i2201
   a1  a2  d1  d2  d3
par 0.966 1.167 3.982 1.714 -0.014
SE  0.053 0.084 0.137 0.087 0.079

$i2202
   a1  a2  d1  d2  d3
par 0.407 0.227 3.917 2.129 1.145
SE  0.048 0.073 0.157 0.091 0.077

$i2203
   a1  a2  d1  d2  d3
par 1.042 0.822 3.437 1.170 -0.465
SE  0.055 0.069 0.128 0.085 0.079

$i2204
   a1  a2  d1  d2  d3
par 1.287 1.056 5.127 1.573 0.525
SE  0.065 0.081 0.187 0.101 0.094

$i2205
   a1  a2  d1  d2  d3
par 0.673 0.776 3.538 1.415 -0.265
SE  0.045 0.067 0.123 0.074 0.068

$i2206
   a1  a2  d1  d2  d3
par 1.155 0.787 3.566 1.820 -0.053
SE  0.074 0.090 0.178 0.125 0.107

$i2207
   a1  a2  d1  d2  d3
par 1.753 1.947 5.000 2.913 1.490
SE  0.136 0.190 0.275 0.220 0.193

$i2208
   a1  a2  d1  d2  d3
par 1.113 0.966 3.872 2.107 0.152
SE  0.063 0.086 0.160 0.112 0.095

$i2209
   a1  a2  d1  d2  d3
par 1.597 1.490 5.137 2.004 -0.024
SE  0.081 0.094 0.197 0.114 0.098

$i2210
   a1  a2  d1  d2  d3
par 1.27 1.413 3.935 1.999 0.560
SE  0.10 0.139 0.192 0.151 0.139

$i2211
   a1  a2  d1  d2  d3
par 0.924 1.027 2.395 0.672 -2.205
SE  0.056 0.090 0.109 0.084 0.108

$i2212
   a1  a2  d1  d2  d3
par 1.685 1.558 3.794 1.793 -0.648
SE  0.076 0.092 0.149 0.117 0.099
### STATISTICAL MODELS OF CHILDREN’S STRATEGY CHANGE IN ANALOGICAL REASONING

<table>
<thead>
<tr>
<th>$i1403$</th>
<th>$a1$</th>
<th>$a2$</th>
<th>$d1$</th>
<th>$d2$</th>
<th>$d3$</th>
<th>$\text{par}$</th>
<th>$1.527$</th>
<th>$0$</th>
<th>$2.545$</th>
<th>$1.966$</th>
<th>$-2.056$</th>
<th>$\text{SE}$</th>
<th>$0.168$</th>
<th>$\text{NA}$</th>
<th>$0.266$</th>
<th>$0.228$</th>
<th>$0.229$</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>$i1404$</th>
<th>$a1$</th>
<th>$a2$</th>
<th>$d1$</th>
<th>$d2$</th>
<th>$d3$</th>
<th>$\text{par}$</th>
<th>$1.704$</th>
<th>$0$</th>
<th>$3.543$</th>
<th>$1.927$</th>
<th>$-1.216$</th>
<th>$\text{SE}$</th>
<th>$0.070$</th>
<th>$\text{NA}$</th>
<th>$0.130$</th>
<th>$0.110$</th>
<th>$0.096$</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>$i1405$</th>
<th>$a1$</th>
<th>$a2$</th>
<th>$d1$</th>
<th>$d2$</th>
<th>$d3$</th>
<th>$\text{par}$</th>
<th>$1.411$</th>
<th>$0$</th>
<th>$3.679$</th>
<th>$2.493$</th>
<th>$-0.563$</th>
<th>$\text{SE}$</th>
<th>$0.196$</th>
<th>$\text{NA}$</th>
<th>$0.330$</th>
<th>$0.292$</th>
<th>$0.257$</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>$i1406$</th>
<th>$a1$</th>
<th>$a2$</th>
<th>$d1$</th>
<th>$d2$</th>
<th>$d3$</th>
<th>$\text{par}$</th>
<th>$1.363$</th>
<th>$0$</th>
<th>$2.990$</th>
<th>$1.123$</th>
<th>$-1.505$</th>
<th>$\text{SE}$</th>
<th>$0.062$</th>
<th>$\text{NA}$</th>
<th>$0.122$</th>
<th>$0.096$</th>
<th>$0.100$</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>$i1501$</th>
<th>$a1$</th>
<th>$a2$</th>
<th>$d1$</th>
<th>$d2$</th>
<th>$d3$</th>
<th>$\text{par}$</th>
<th>$1.908$</th>
<th>$0$</th>
<th>$3.248$</th>
<th>$1.738$</th>
<th>$-0.502$</th>
<th>$\text{SE}$</th>
<th>$0.149$</th>
<th>$\text{NA}$</th>
<th>$0.223$</th>
<th>$0.188$</th>
<th>$0.164$</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>$i1502$</th>
<th>$a1$</th>
<th>$a2$</th>
<th>$d1$</th>
<th>$d2$</th>
<th>$d3$</th>
<th>$\text{par}$</th>
<th>$2.077$</th>
<th>$0$</th>
<th>$2.790$</th>
<th>$1.236$</th>
<th>$-2.153$</th>
<th>$\text{SE}$</th>
<th>$0.091$</th>
<th>$\text{NA}$</th>
<th>$0.128$</th>
<th>$0.114$</th>
<th>$0.118$</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>$i1503$</th>
<th>$a1$</th>
<th>$a2$</th>
<th>$d1$</th>
<th>$d2$</th>
<th>$d3$</th>
<th>$\text{par}$</th>
<th>$1.540$</th>
<th>$0$</th>
<th>$3.911$</th>
<th>$2.396$</th>
<th>$-1.037$</th>
<th>$\text{SE}$</th>
<th>$0.072$</th>
<th>$\text{NA}$</th>
<th>$0.132$</th>
<th>$0.108$</th>
<th>$0.098$</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>$i1504$</th>
<th>$a1$</th>
<th>$a2$</th>
<th>$d1$</th>
<th>$d2$</th>
<th>$d3$</th>
<th>$\text{par}$</th>
<th>$1.987$</th>
<th>$0$</th>
<th>$2.792$</th>
<th>$1.187$</th>
<th>$-2.030$</th>
<th>$\text{SE}$</th>
<th>$0.084$</th>
<th>$\text{NA}$</th>
<th>$0.117$</th>
<th>$0.106$</th>
<th>$0.118$</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>$i1505$</th>
<th>$a1$</th>
<th>$a2$</th>
<th>$d1$</th>
<th>$d2$</th>
<th>$d3$</th>
<th>$\text{par}$</th>
<th>$1.177$</th>
<th>$0$</th>
<th>$3.809$</th>
<th>$2.387$</th>
<th>$-0.096$</th>
<th>$\text{SE}$</th>
<th>$0.160$</th>
<th>$\text{NA}$</th>
<th>$0.426$</th>
<th>$0.249$</th>
<th>$0.181$</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>$i1601$</th>
<th>$a1$</th>
<th>$a2$</th>
<th>$d1$</th>
<th>$d2$</th>
<th>$d3$</th>
<th>$\text{par}$</th>
<th>$1.320$</th>
<th>$0$</th>
<th>$2.019$</th>
<th>$0.955$</th>
<th>$-1.966$</th>
<th>$\text{SE}$</th>
<th>$0.058$</th>
<th>$\text{NA}$</th>
<th>$0.094$</th>
<th>$0.086$</th>
<th>$0.095$</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>$i1602$</th>
<th>$a1$</th>
<th>$a2$</th>
<th>$d1$</th>
<th>$d2$</th>
<th>$d3$</th>
<th>$\text{par}$</th>
<th>$1.662$</th>
<th>$0$</th>
<th>$1.992$</th>
<th>$0.775$</th>
<th>$-2.254$</th>
<th>$\text{SE}$</th>
<th>$0.069$</th>
<th>$\text{NA}$</th>
<th>$0.101$</th>
<th>$0.096$</th>
<th>$0.115$</th>
</tr>
</thead>
</table>

<p>| $i1603$ | $a1$ | $a2$ | $d1$ | $d2$ | $d3$ | $\text{par}$ | $1.736$ | $0$ | $2.802$ | $1.312$ | $-1.702$ | $\text{SE}$ | $0.146$ | $\text{NA}$ | $0.215$ | $0.187$ | $0.185$ |</p>
<table>
<thead>
<tr>
<th>( a_1 )</th>
<th>( a_2 )</th>
<th>( d_1 )</th>
<th>( d_2 )</th>
<th>( d_3 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \text{par} )</td>
<td>( 2.039 )</td>
<td>0</td>
<td>3.367</td>
<td>1.145</td>
</tr>
<tr>
<td>( \text{SE} )</td>
<td>0.437</td>
<td>NA</td>
<td>0.499</td>
<td>0.370</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>( a_1 )</th>
<th>( a_2 )</th>
<th>( d_1 )</th>
<th>( d_2 )</th>
<th>( d_3 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \text{par} )</td>
<td>1.696</td>
<td>0</td>
<td>2.728</td>
<td>1.334</td>
</tr>
<tr>
<td>( \text{SE} )</td>
<td>0.080</td>
<td>NA</td>
<td>0.122</td>
<td>0.108</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>( a_1 )</th>
<th>( a_2 )</th>
<th>( d_1 )</th>
<th>( d_2 )</th>
<th>( d_3 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \text{par} )</td>
<td>1.264</td>
<td>0</td>
<td>1.160</td>
<td>0.645</td>
</tr>
<tr>
<td>( \text{SE} )</td>
<td>0.117</td>
<td>NA</td>
<td>0.143</td>
<td>0.135</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>( a_1 )</th>
<th>( a_2 )</th>
<th>( d_1 )</th>
<th>( d_2 )</th>
<th>( d_3 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \text{par} )</td>
<td>1.250</td>
<td>0</td>
<td>1.626</td>
<td>0.450</td>
</tr>
<tr>
<td>( \text{SE} )</td>
<td>0.073</td>
<td>NA</td>
<td>0.121</td>
<td>0.106</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>( a_1 )</th>
<th>( a_2 )</th>
<th>( d_1 )</th>
<th>( d_2 )</th>
<th>( d_3 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \text{par} )</td>
<td>2.364</td>
<td>0</td>
<td>2.574</td>
<td>1.114</td>
</tr>
<tr>
<td>( \text{SE} )</td>
<td>0.474</td>
<td>NA</td>
<td>0.474</td>
<td>0.401</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>( a_1 )</th>
<th>( a_2 )</th>
<th>( d_1 )</th>
<th>( d_2 )</th>
<th>( d_3 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \text{par} )</td>
<td>3.104</td>
<td>0</td>
<td>4.111</td>
<td>2.683</td>
</tr>
<tr>
<td>( \text{SE} )</td>
<td>0.615</td>
<td>NA</td>
<td>0.672</td>
<td>0.580</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>( a_1 )</th>
<th>( a_2 )</th>
<th>( d_1 )</th>
<th>( d_2 )</th>
<th>( d_3 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \text{par} )</td>
<td>1.423</td>
<td>0</td>
<td>0.850</td>
<td>0.654</td>
</tr>
<tr>
<td>( \text{SE} )</td>
<td>0.172</td>
<td>NA</td>
<td>0.184</td>
<td>0.181</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>( a_1 )</th>
<th>( a_2 )</th>
<th>( d_1 )</th>
<th>( d_2 )</th>
<th>( d_3 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \text{par} )</td>
<td>1.276</td>
<td>0</td>
<td>1.800</td>
<td>0.550</td>
</tr>
<tr>
<td>( \text{SE} )</td>
<td>0.073</td>
<td>NA</td>
<td>0.121</td>
<td>0.106</td>
</tr>
</tbody>
</table>
4.2 M1 (M0 + explanatory person predictor training condition)

M1 <- mixedmirt(data=DATA, covdata, mirt.model(S0), itemtype='graded',
      technical = list(NCYCLES = 1e5), lr.fixed = list(F1 = ~ e20092 + e20101 +
      e20102 + e20111 + e20121, F2 = ~ e20092 + e20101 + e20102 + e20111 + e20121 +
      Control + Practice + Feedback))

> M1
Full-information item factor analysis with 2 factor(s).
Converged within 0.001 tolerance after 936 MHRM iterations.
mirt version: 1.8
M-step optimizer: NR

Information matrix estimated with method: MHRM
Condition number of information matrix = 1778.582
Second-order test: model is a possible local maximum
Log-likelihood = -38190.69, SE = 0.206
AIC = 76731.37; AICc = 76805.5
BIC = 77591.45; SABIC = 77035.64

> summary(M1)
----------
RANDOM EFFECT COVARIANCE(S):
Correlations on upper diagonal
$\text{Theta}$

\begin{array}{ll}
F1 & 1.000 \\
F2 & -0.387
\end{array}

\begin{array}{l}
F1 \\
F2
\end{array}

\begin{array}{l}
-0.387 \\
1.000
\end{array}

\begin{array}{l}
\text{LATENT REGRESSION FIXED EFFECTS:}
\end{array}

\begin{array}{llllll}
(\text{Intercept}) & 0.000 & 0.000 \\
e20092 & -1.231 & 1.116 \\
e20101 & -0.526 & 0.944 \\
e20102 & -1.248 & 1.189 \\
e20111 & -1.197 & 1.632 \\
e20121 & 0.634 & 1.465 \\
\text{Control} & 0.000 & -0.327 \\
\text{Practice} & 0.000 & -0.610 \\
\text{Feedback} & 0.000 & -0.685
\end{array}

\begin{array}{llllll}
\text{Std.Error}_{\text{F1}} & 0.125 & 0.347 & -9.852 & 3.219 \\
\text{Std.Error}_{\text{F2}} & 0.138 & 0.219 & -3.806 & 4.309 \\
e20092 & 0.062 & 0.076 & -20.146 & 15.656 \\
e20101 & 0.062 & 0.074 & -19.221 & 21.974 \\
e20102 & 0.059 & 0.089 & 10.830 & 16.470 \\
\text{Control} & NA & 0.106 & NA & -3.078 \\
\text{Practice} & NA & 0.082 & NA & -7.437 \\
\text{Feedback} & NA & 0.305 & NA & -2.247
\end{array}

4.3 M2 (M1 + explanatory person predictor age)

\begin{verbatim}
M2 <- mixedmirt(data=DATA, covdata, mirt.model(S0), itemtype='graded',
technical = list(NCYCLES = 1e5), lr.fixed = list(F1 = ~ e20092 + e20101 +
e20102 + e20111 + e20121 + age_cent, F2 = ~ e20092 + e20101 + e20102 +
e20111 + e20121 + Control + Practice + Feedback + age_cent))
\end{verbatim}

> M2

Full-information item factor analysis with 2 factor(s).
Converged within 0.001 tolerance after 705 MHRM iterations.
mirt version: 1.8
M-step optimizer: NR

Information matrix estimated with method: MHRM
Condition number of information matrix = 3866.075
Second-order test: model is a possible local maximum

Log-likelihood = -38049.9, SE = 0.208
AIC = 76453.8; AICc = 76529.81
BIC = 77323.7; SABIC = 76761.54
> summary(M2)

----------
RANDOM EFFECT COVARIANCE(S):
Correlations on upper diagonal

$Theta
F1  F2
F1  1.000 -0.356
F2  -0.356  1.000

----------
LATENT REGRESSION FIXED EFFECTS:

F1  F2
  (Intercept) 0.000  0.000
  e20092  -0.185  0.615
  e20101  -0.551  0.897
  e20102  -1.294  1.145
  e20111  -1.296  1.633
  e20121  -1.040  2.179
  age_cent  0.685  -0.279
  Control    0.000  -0.346
  Practice   0.000  -0.645
  Feedback   0.000  -0.702

Std.Error_F1 Std.Error_F2   z_F1   z_F2
  (Intercept) NA       NA     NA     NA
  e20092  0.136       0.347 -1.360  1.772
  e20101  0.139       0.220 -3.968  4.079
  e20102  0.062       0.071-20.769 16.132
  e20111  0.062       0.072-21.077 22.567
  e20121  0.099       0.128-10.492 17.083
  age_cent 0.032       0.038 21.311 -7.269
  Control   NA       0.102  NA -3.402
  Practice  NA       0.083  NA  -7.763
  Feedback  NA       0.305  NA -2.302

4.4 M3 (M2 + interaction training condition*age)

M3 <- mixedmirt(data=DATA, covdata, mirt.model(S0), itemtype='graded',
technical = list(NCYCLES = 1e5), lr.fixed = list(F1 = ~ e20092 + e20101 +
e20102 + e20111 + e20121 + age_cent, F2 = ~ e20092 + e20101 + e20102 +
e20111 + e20121 + Control + Practice + Feedback + age_cent +
Control*age_cent + Practice*age_cent + Feedback*age_cent))

> M3

Full-information item factor analysis with 2 factor(s).
Converged within 0.001 tolerance after 873 MHRM iterations.
mirt version: 1.8
M-step optimizer: NR

Information matrix estimated with method: MHRM
Condition number of information matrix = 4066.54
Second-order test: model is a possible local maximum

Log-likelihood = -38038.71, SE = 0.207
AIC = 76437.42; AICc = 76516.3
BIC = 77322.07; SABIC = 76750.37

> summary(M3)

RANDOM EFFECT COVARIANCE(S):
Correlations on upper diagonal

$Theta

<table>
<thead>
<tr>
<th></th>
<th>F1</th>
<th>F2</th>
</tr>
</thead>
<tbody>
<tr>
<td>F1</td>
<td>1.000</td>
<td>-0.344</td>
</tr>
<tr>
<td>F2</td>
<td>-0.344</td>
<td>1.000</td>
</tr>
</tbody>
</table>

LATENT REGRESSION FIXED EFFECTS:

<table>
<thead>
<tr>
<th></th>
<th>F1</th>
<th>F2</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>e20092</td>
<td>-0.192</td>
<td>1.443</td>
</tr>
<tr>
<td>e20101</td>
<td>-0.577</td>
<td>0.904</td>
</tr>
<tr>
<td>e20102</td>
<td>-1.313</td>
<td>1.142</td>
</tr>
<tr>
<td>e20111</td>
<td>-1.317</td>
<td>1.802</td>
</tr>
<tr>
<td>e20121</td>
<td>-1.076</td>
<td>2.184</td>
</tr>
<tr>
<td>age_cent</td>
<td>0.691</td>
<td>-0.367</td>
</tr>
<tr>
<td>Control</td>
<td>0.000</td>
<td>-0.600</td>
</tr>
<tr>
<td>Practice</td>
<td>0.000</td>
<td>-0.638</td>
</tr>
<tr>
<td>Feedback</td>
<td>0.000</td>
<td>-0.570</td>
</tr>
<tr>
<td>Control:age_cent</td>
<td>0.000</td>
<td>0.322</td>
</tr>
<tr>
<td>Practice:age_cent</td>
<td>0.000</td>
<td>0.058</td>
</tr>
<tr>
<td>Feedback:age_cent</td>
<td>0.000</td>
<td>0.586</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Std.Error_F1</th>
<th>Std.Error_F2</th>
<th>z_F1</th>
<th>z_F2</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>e20092</td>
<td>0.135</td>
<td>0.570</td>
<td>-1.423</td>
<td>2.533</td>
</tr>
<tr>
<td>e20101</td>
<td>0.138</td>
<td>0.220</td>
<td>-4.173</td>
<td>4.116</td>
</tr>
<tr>
<td>e20102</td>
<td>0.066</td>
<td>0.077</td>
<td>-19.919</td>
<td>14.801</td>
</tr>
<tr>
<td>e20111</td>
<td>0.068</td>
<td>0.089</td>
<td>-19.293</td>
<td>20.212</td>
</tr>
<tr>
<td>e20121</td>
<td>0.087</td>
<td>0.108</td>
<td>-12.365</td>
<td>20.161</td>
</tr>
<tr>
<td>age_cent</td>
<td>0.037</td>
<td>0.046</td>
<td>18.643</td>
<td>-8.014</td>
</tr>
<tr>
<td>Control</td>
<td>NA</td>
<td>0.114</td>
<td>NA</td>
<td>-5.277</td>
</tr>
<tr>
<td>Practice</td>
<td>NA</td>
<td>0.080</td>
<td>NA</td>
<td>-7.933</td>
</tr>
<tr>
<td>Feedback</td>
<td>NA</td>
<td>0.314</td>
<td>NA</td>
<td>-1.815</td>
</tr>
<tr>
<td>Control:age_cent</td>
<td>NA</td>
<td>0.068</td>
<td>NA</td>
<td>4.733</td>
</tr>
<tr>
<td>Practice:age_cent</td>
<td>NA</td>
<td>0.057</td>
<td>NA</td>
<td>1.010</td>
</tr>
<tr>
<td>Feedback:age_cent</td>
<td>NA</td>
<td>0.274</td>
<td>NA</td>
<td>2.137</td>
</tr>
</tbody>
</table>