Puzzling with potential : dynamic testing of analogical reasoning in children

Stevenson, C.E.

Author: Stevenson, Claire Elisabeth
Title: Puzzling with potential : dynamic testing of analogical reasoning in children
Issue Date: 2012-09-13

CHAPTER 5

Explanatory item response modeling of children's change on a dynamic test of analogical reasoning

This chapter is based on Stevenson, C. E., Hickendorff, M., Heiser, W. J., Resing, W. C. M. & De Boeck, P. A. L. (under review). Explanatory item response modeling of children's change on a dynamic test of analogical reasoning.

Abstract

Dynamic testing is an assessment method in which training is incorporated into the testing procedure with the aim of gauging cognitive potential. Large individual differences are present in children's ability to profit from training in analogical reasoning. The aim was to investigate sources of these differences on a dynamic test of figural analogies. School children (N=252, M=7 years, SD=11 months, range 5-9 years) were dynamically tested using a pretest-training-posttest design. The children were randomly allocated to a training condition: graduated prompts or feedback. All children were presented with figural analogies without help or feedback during the pretest. The children then received training on the analogy task. This was followed by the posttest measure. Explanatory IRT models were used to investigate sources of individual differences in initial ability and improvement after training. We found that visual and verbal working memory and age were related to initial ability. Improvement after training was influenced by training-type, whereby graduated prompts trained children improved more than feedback-trained, but also by initial ability, where children with lower initial scores improved more in both conditions. Furthermore, degree of improvement was related to math achievement; where higher achieving children improved more from pretest to posttest. Potential to learn as measured by dynamic tests is not often included in traditional cognitive assessment. However, learning potential does appear to be an important construct to include in psychoeducational testing.

Acknowledgments

We would like to thank Carlijn Bergwerff for organizing the data collection and Aafke Snelting, Bart Leenhouts, Isabelle Neerhout, Janneke de Ruiter, Margreet van Volkom, Marit Ruijgrok, Nienke Faber, Noraly Snel and Rosa Alberto for their assistance with data collection and coding.

5.1 Introduction

Dynamic testing can be seen as an assessment form that aims to tap into the test taker's potential for learning by assessing what can be learned over a short period of time in which instruction in problem solving is provided (Elliott, 2003; Sternberg & Grigorenko, 2002). The main difference between dynamic and traditional assessment methods is that dynamic testing incorporates feedback into the assessment process (Elliott et al., 2010; Grigorenko & Sternberg, 1998). Dynamic testing is often contrasted with traditional "static" testing such as administering an IQ test in which no feedback or training is given. In some situations, static tests provide a sound indication of a person's present capabilities and predict academic success or failure (e.g., Neisser et al., 1996; Sternberg et al., 2001). Researchers and educational practitioners agree that an indication of a child's potential for learning could contribute to psychoeducational assessment (Elliott et al., 2010; Jeltova et al., 2007). Dynamic tests can provide information on learning potential through indices such as gain scores (improvement from pretest to posttest), instructional-needs (e.g., Bosma & Resing, 2012; Jeltova et al., 2011) or strategy development (e.g., Resing & Elliott, 2011; Resing et al., 2009). A major obstacle within the field of dynamic testing however has been how to obtain and interpret reliable measures of individual differences in cognitive potential (Embretson, 1991b; Sternberg & Grigorenko, 2002). Item response theory (e.g., Rasch, 1961), potentially offers ways to solve the inherent problems of measuring learning and change (e.g., Embretson, 1991b, 1991a). Aim of the present study was to extend item response modeling of dynamic testing performance not only to measure individual differences in children's cognitive potential but also to explain the differences in training effects in terms of variations in age, working memory and previous school performance using explanatory item response theory (IRT) (De Boeck & Wilson, 2004).

5.1.1 Individual differences in cognitive potential

The ability to learn can be considered one of the many constructs that falls under the term intelligence (e.g., Sternberg & Kaufmann, 2011; Neisser et al., 1996), and individual differences in the ability to learn may form a dynamic component of this concept. Recent research seems to indicate that fluid reasoning ability may be more influenced by learning experiences than thought before. For example, there appear to be considerable individual differences in the effects of retesting and training on fluid reasoning tasks in both adults (Freund & Holling, 2011a) and school children (Freund & Holling, 2011b; Mackey, Hill, Stone, & Bunge, 2010). Working memory training also appears to influence performance in the short-term on tests of fluid reasoning in adults (Jaeggi et al., 2008) and preschoolers (Thorell, Lindqvist, Nutley, S Bohlin, & Klingberg, 2009). These findings on the modifiability of cognitive capacities can be interpreted within the theoretical framework of dynamic testing – where abilities are considered flexible rather than fixed in a context of developing expertise (Grigorenko & Sternberg, 1998; Sternberg & Grigorenko, 2002). Similarly, the results of dynamic testing studies, which often comprise of a pretest-training-posttest design, coincide with research on retesting and training effects of fluid intelligence as generally positive training effects are found, interestingly again with large individual variation in improvement (e.g., Fabio, 2005; Jeltova et al., 2011; Swanson & Lussier, 2001; Sternberg et al., 2007).

The idea behind dynamic testing is that a traditionally administered standardized test measures one's present capacities, whereas dynamic testing may provide information about one's potential for learning. This information may be of additional value to static test results in the prediction of scholastic achievement (e.g., Caffrey et al., 2008; L. S. Fuchs et al., 2008; Hessels, 2009; Resing, 1997; Stevenson, Heiser, & Resing, submitted 2012b) and provision of information to help improve school performance (e.g., Bosma & Resing, 2012; Bosma et al., submitted; Jeltova et al., 2007, 2011; Grigorenko, 2009a).

5.1.2 Measuring Learning Potential with Dynamic Testing

Whereas in static tests, provision of feedback is often viewed as a source of error, in dynamic testing the ability to profit from training is considered a way of uncovering potential cognitive capacity (Embretson, 1991b; Embretson & Prenovorst, 2000; Sternberg & Grigorenko, 2002). In the typical dynamic testing pretest-training-posttest design, structured feedback is provided during one or more training sessions. Presently, posttest scores are most often used as an indication of children's potential ability because gain scores (posttest minus pretest score) may be unreliable in the context of classical test theory (Resing, Elliott, & Grigorenko, 2012). Using raw gain scores to measure change leads to various problems (e.g., De Bock, 1976; Embretson, 1991b), such as the unreliability of the gain score, the fact that the scale units for change do not have a constant meaning for test takers with different pretest scores and the regression effect of repeated administration (Lord, 1963). These problems are potentially solved when IRT is employed because the ability scores for pretest and posttest are no longer ordinal measures, but are put on a joint interval measurement scale using logistic models (Embretson & Reise, 2000). In the simplest IRT model, the Rasch model, the chance that an item is solved correctly depends on the difference between the latent ability of the examinee and the difficulty of the item. Here the IRT Rasch-based change score has the same meaning across the whole range of the measurement scale in terms of log odds (i.e. the logarithm of probability of correct vs. incorrect). Thus IRT is appropriate for measuring change as it provides a good basis for the latent scaling of gain scores and problems with unreliability are dealt with as reliability is separated from other parts of the model (Embretson & Reise, 2000).

In the dynamic assessment literature, classical test measures tend to dominate (e.g., Calero et al., 2011; Resing, Steijn, Xenidou-Dervou, Stevenson, & Elliott, 2011; Tzuriel & Egozi, 2010). Earlier findings based on classical test theory may still hold if pretest-posttest control group designs are used, provided there are few pretest-differences between the groups and there are no floor or ceiling effects for either of the groups. However, the focus of dynamic testing is not only on the measurement of the average gain from training, but rather on identifying how and why some children profit more from training than others – i.e. individual differences in learning and change (e.g., Resing & Elliott, 2011; Resing et al., 2009) – so that timely intervention can be provided (Caffrey et al., 2008; Elliott, 2003). In an educational setting the assumption is that there are individual differences both in initial ability and ability to profit from instruction. It is therefore imperative to have good gain estimates when investigating the sources of these differences in individual change. IRT models seem appropriate for this purpose.

IRT measurement models for dynamic tests have gained some ground. For example in the Hessel's Analogical Reasoning Test (HART) with a train-test format used Rasch scaling of the test session (Hessels & Bosson, 2003). De Beer also used Rasch item calibration for her computer adaptive test of Learning Potential (De Beer, 2005). Embretson (1991b) developed the Multidimensional Rasch Model for Learning and Change (MRMLC) to measure ability and modifiability (i.e. performance change) from one testing occasion to the next and applied this to a dynamic test of visuospatial reasoning (Embretson, 1987, 1992). In research with AnimaLogica, the dynamic test of figural analogical reasoning employed in the present study, we have also applied MRMLC to measure pretest ability and performance change after training 3. These are examples of IRT being used purely for measurement purposes. However, IRT can also be used as a research tool – for example to investigate cognitive processes (e.g., De Boeck, Wilson, & Acton, 2005) or explain learning in developmental psychology (e.g., Janssen, De Boeck, Viane, & Vallaeys, 1999) and educational psychology (e.g., Hickendorff, Van Putten, Verhelst, & Heiser, 2010). With IRT it is possible to combine both measurement and explanation of individual differences and item effects in one and the same analysis – a method De Boeck and Wilson (2004) coined as explanatory IRT– which we applied in the present study to measure and explain children's ability and potential on an dynamically administered analogical reasoning task.

5.1.3 Dynamic testing of analogical reasoning

This article focuses on explaining individual differences in children's performance on a dynamic test of analogical reasoning by investigating combinations of explanatory variables using IRT models to estimate the change in ability. We examined the combined contribution of variables previously implicated as related to children's progression in analogy solving: (1) training-type, (2) age, (3) working memory capacity, (4) initial ability and (5) school performance.

In the current study we used figural matrix analogies (see Figure 5.1), which are a classical form of analogies (A:B::C:?) often utilized in psychoeducational assessment to measure fluid reasoning capacity, such as the Raven Standard Progressive Matrices (Raven, Raven, & Court, 2004). Performance on matrix analogies has been found to be related to school performance (Balboni et al., 2010; Ferrer & McArdle, 2004; Hessels, 2009) – especially math achievement (Primi, Eugénia Ferrao, & Almeida, 2010; Taub, Floyd, Keith, & McGrew, 2008) – and is considered an important ability required in school learning (Goswami, 1992).

On the whole, older children generally solve analogy problems better than younger children (e.g., Csapó, 1997; Hosenfeld & Resing, 1997; Sternberg & Rifkin, 1979). In Siegler & Svetina's (2002) microgenetic and cross-sectional study of children's analogical reasoning initially six year-olds solve significantly fewer analogies than the older children included in the study. However, after repeated practice the six year-olds on average perform at a similar level as seven and eight year-olds. Yet, children's ability to solve figural analogies appears to develop with great variability throughout childhood evidenced by large differences within each age group both in initial ability as well as performance change (e.g., Cheshire et al., 2005; Siegler & Svetina, 2002; Stevenson et al., 2011, under review; Tunteler et al., 2008).

Working memory efficiency also shows developmental increases with age, and is a well-researched source of individual differences in fluid reasoning in children (e.g., Alloway et al., 2004; Engel de Abreu et al., 2010; Tillman et al., 2008). Improvement in working memory (WM) seems to correspond with improvement in reasoning and problem solving in children (Fry & Hale, 1996; Kail, 2007; Swanson, 2008). Children's ability to solve figural analogies appears to be related to their working memory efficiency (e.g., Richland et al., 2006; Tunteler & Resing, 2010). For example, both verbal and visuospatial components were found to coincide with children's performance on tests with figural matrices (Hornung et al., 2011; Stevenson et al., submitted 2011a). Therefore measures of both visuospatial and verbal working memory were included as possible sources of individual differences in initial ability and performance change in the present study.

The type of training provided in a test-train-test design can be a source of individual differences in change (Ball, Hoyle, & Towse, 2010; Harpaz-Itay et al., 2006; Stevenson et al., under review; Tunteler et al., 2008). For example, Resing et al., (2009) found that the graduated prompts method, a specific form of training providing increasingly elaborate instructions of metacognitive skills, cognitive processing components and task-specific scaffolds on solution strategies, led to different paths of strategy-change in Dutch and ethnic minority children. Luwel, Foustana, Papadatos & Verschaffel (2010) demonstrated that strategy feedback training improved low IQ children's numerosity judgment task performance more so than outcome feedback, but high IQ children's improvement was not moderated by training-type. The literature generally seems to indicate that children with lower initial ability tend to improve more during dynamic testing (Swanson & Lussier, 2001). Although, in some cases it is possible that this is due to ceiling effects (Sternberg & Grigorenko, 2002). We chose to use moderately difficult items in our dynamic test and IRT to model performance change in order to avoid this problem. In the present study we investigated whether graduated prompts training versus outcome feedback training led to differential changes in figural analogy solving and whether this interacts with age, working memory, initial ability or school performance to explain individual differences in change.

5.1.4 Current Study

The present study aimed to explain children's differences in change in analogical reasoning skills using the explanatory IRT framework. Our first research question concerned whether children's performance, as a consequence of training would (1a) progress from pretest to posttest, and (1b) show individual differences in degree of improvement (e.g., Embretson, 1987; Freund & Holling, 2011a, 2011b). Our second research question focused on the effect of type of training. We expected (2a) the children in the graduated prompts condition would progress more on average in analogy solving than children who received outcome feedback (e.g., Luwel et al., 2010). Furthermore, we hypothesized (2b) that children with lower initial ability would generally improve more than those with higher initial ability (e.g., Luwel et al., 2010; Swanson & Lussier, 2001). Our third research question concerned whether the children's performance and progress was best explained by age, working memory or by a combination of these variables. We expected (3a) that older children would perform better on the analogies than younger, less experienced peers (e.g., Siegler & Svetina, 2002) and (3b) that children with greater WM efficiency would on average display greater proficiency in analogical reasoning (e.g., Richland et al., 2006; Stevenson et al., submitted 2011a). Next, we examined whether (3c) WM capacity or (3d) age interacted with the children's ability to profit from training. Finally given the relationship of matrix analogy solving with mathematics (e.g., Primi et al., 2010), we investigated (4) if school performance was also related to the children's performance change from pretest to posttest.

5.2 Method

5.2.1 *Sample*

255 children from three age-groups (kindergarten, first and second grade) were recruited from five intercity public elementary schools of similar middle class SES in the south-west of the Netherlands. The sample consisted of 119 boys and 136 girls, with a mean age of 7 years, 11 months (range 4;11-9;3 years). The schools were selected based on their willingness to participate. Written informed consent for children's participation was obtained from the parents.

5.2.2 *Design & Procedure*

A pretest-training-posttest control-group design with randomized blocking was employed. Children were randomly assigned to a training-type condition: (1) graduated prompts or (2) outcome feedback, based on school, classroom, gender and age. Sessions took place weekly and all participants were tested individually in a quiet room at the child's school by educational psychology students trained in the procedure. Each session lasted approximately 20 minutes and total testing time comprised less than 1.5 hours. During the first session, all participants were administered the working memory tasks, a computer mouse task, and the AnimaLogica analogies-introduction task. The computer mouse task (Stevenson et al., 2011) was administered prior to testing to ensure that the children were able to perform the necessary clicking and drag & drop actions required for the dynamic analogy test. An analogies-introduction task (see Stevenson et al., 2009), based on the objects and transformations used in the analogy task, was also administered to ensure that the children were familiar with the content prior to testing.

The AnimaLogica pretest was administered during the second session. The two following sessions comprised of training in analogy solving. Half of the children were trained according to the graduated prompts method and the other half received outcome feedback training (described in section 2.3). The posttest was administered during the final session. All instructions were provided according to standardized protocols (see 3).

5.2.3 Measures

AnimaLogica*: a dynamic test of figural analogical reasoning*

AnimaLogica is a computerized dynamic test of analogical reasoning for children. The figural analogies (A:B::C:?) comprised of 2x2 matrices with familiar animals as objects (see Figure 5.1). The animals changed horizontally or vertically by color, orientation, size, position, quantity or animal type. The number of transformations – or object changes – were used to gauge item difficulty (e.g., Hosenfeld & Resing, 1997; Mulholland et al., 1980). The items difficulties ranged from two transformations to eight transformations. The children had to construct the solution using a computer mouse to drag & drop animal figures representing the six transformations into the empty box in the lower left or right quadrant of the matrix. A maximum of two animals were present in each analogy. These were available in three colors (red, yellow, blue) and two sizes (large, small). The orientation (facing left or right) could be changed by clicking the figure. Quantity was specified by the number of figures placed in the empty box. Position was specified by location of the figure placed in the box.

Pretest and Posttest. The test booklets consisted of 20 items of varied difficulty. The pretest and posttest items were isomorphs (e.g., Freund & Holling, 2011a) in which the items only differed in color and type of animal, but the exact same transformations were used. Given the young sample, items with 2-4 transformations were emphasized in test construction. More specifically, the difficulty level (based on number of transformations) of the pretest and posttest items was as follows: four items of difficulty levels 2 to 4, three items of difficulty levels 5 and 6 and one item each for difficulty levels 7 & 8. The items were then randomly selected from a pool of possible items using constraints that allowed for a balanced representation of each of the animals, colors and transformations in the test.

Training. The training consisted of the same figural analogy matrices. The 10 training items did not occur in the tests. Two training methods were applied: graduated prompts or outcome feedback. The graduated prompts method (e.g., Campione & Brown, 1987; Resing, 1997; Resing & Elliott, 2011; Resing et al., 2009; Stevenson et al., under review, submitted 2011a) consisted of stepwise instructions and began with general, metacognitive prompts, such as focusing attention, followed by cognitive hints, emphasizing the transformations and solution procedure, and ended with step-by-step scaffolds to solve the problem. A maximum of five prompts were administered. Once the child answered an item correctly the child was asked to explain his/her answer; no further prompts were provided and the examiner proceeded with the next item. Outcome feedback training also allowed for 4 attempts to correctly solve each item. However, the children were only told if their solution was correct or incorrect and received motivational comments. After a correct solution or 4 attempts no further feedback was given and the examiner proceeded with the next item.

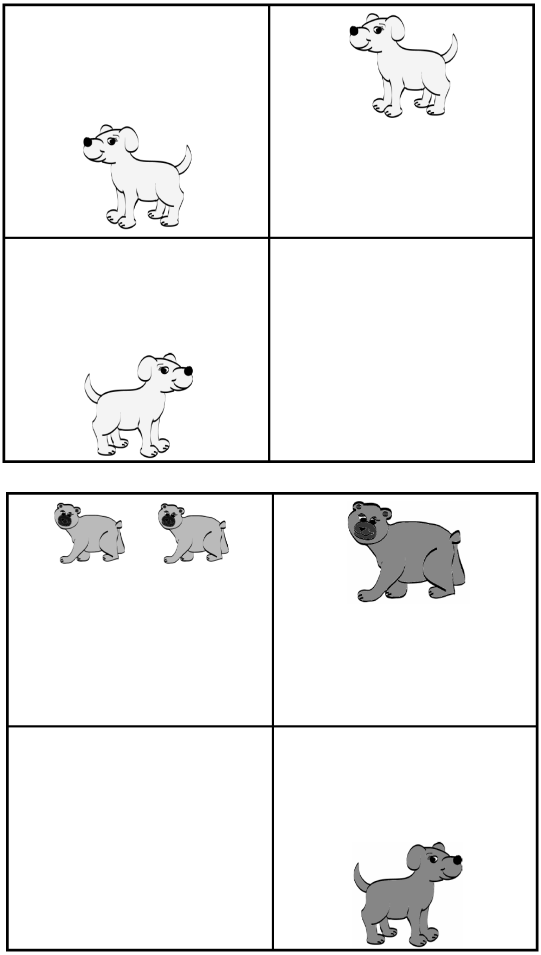

Figure 5.1 *Examples of figural matrix analogies used in* AnimaLogica. *Top figure contains two transformations (horizontal: position; vertical: orientation). Bottom figure contains six transformations (horizontal: color, quantity and size; vertical: animal, orientation and position).*

Automated Working Memory Assessment (AWMA, Alloway, 2007)

Listening Recall. This verbal working memory subtest consists of spoken sentences, of which the child is asked to repeat the first word and say whether the sentence is true or false (e.g., bicycles can walk).

Spatial Span. In this visuospatial working memory subtest a sequence of two figures are presented and the child is asked to say whether these are the same or different. In some cases one of the figures is rotated (i.e. same) and others mirrored and rotated (i.e. different). The child must also recall in sequence whether the red dots were located above, left or right of the figure on the right.

Math achievement

The children each took part in biannual scholastic achievement assessments administered in the classroom by the child's teacher in January and June of each school year (CITO , 2010a, 2010b, 2010c). These multiple-choice tests are widely used at primary schools in the Netherlands for the purpose of tracking children's performance on school subjects. The math test items are similar for the included age-groups and involve pictorial or number problems mostly concerning number relations, addition and subtraction, but for the second graders also a few geometry or multiplication/division problems (CITO , 2010a, 2010b, 2010c). The scores are based on national norms per age-group and range from A to E; 'A' is categorized as a very good, indicating a performance falling within the top 25 percent. 'B' scores (good) are between 26th and 50th percentile whereas 'C' scores (sufficient) indicate 51st to 75th percentile performance. 'D' (weak) and 'E' (very weak) scores fall within the lowest 25% – 'D' scores indicate performance with the 11th to 25th percentile range and 'E' scores fall in the lowest 10%.

5.3 Results

5.3.1 *Initial Group Comparisons*

The substantive aims of this paper focused on the role training-type, age, working memory and prior school performance (math achievement scores) play in children's analogical reasoning progression in a dynamic testing context. It is therefore important to investigate whether group differences were present prior to dynamic testing. The children in the two training conditions did not differ in age ($t(250) = -.46, p = .65$) or working memory capacity (listening recall: $t(250) = 1.63, p = .11$ or spatial span: $t(250) = .66, p = .51$) and they were equally divided per school year ($\chi^2(3) = .30, p = .96$) and gender ($\chi^2(1) = .05, p = .82$). Age and working memory correlated moderately (listening recall: $r = .44, p < .001$ and spatial span: $r = .48, p < .001$). The children in the three different school years naturally differed in age ($F(3, 248) = 218.92, p < .001$) and working memory scores (listening recall: $F(3, 248) = 36.24, p < .001$ and spatial span: $F(3, 248) = 41.62, p < .001$). The children's median scores on the math achievement test were near the national mean and a Kruskal-Wallis test showed that the distribution of the math achievement scores was similar across the three grades, $\chi^2(2) = 1.50, p = .47$, and two conditions: $\chi^2(1) = 2.69, p = .30$. See Table 5.1 for descriptive statistics.

5.3.2 *Psychometric Properties*

Cronbach's alpha coefficient of internal consistency was $\alpha = .904$ for the pretest and $\alpha = .906$ for the posttest. The reliabilities of the test on both sessions are considered very satisfactory. The pretest proportion correct responses per item ranged from .02 to .60 and for the posttest from .12 to .84. The rank correlation between the proportion incorrect and the predicted difficulty level based on the number of transformations was $\rho = .86, p < .001$ for the pretest and $\rho = .86, p < .001$

Table 5.1 *Means and standard deviations of age and working memory scores per age-group and condition (GP=graduated prompts, FB=feedback).*

			Age		Listening Recall		Spatial Span		Math Achievement		
Age group		N	M	SD	M	SD	M	SD	M	Median	SD
Kindergarten	GP	38	70.11	4.40	6.70	3.11	8.16	5.95	3.30	3.00	1.42
	FB	37	70.16	4.63	6.84	3.53	9.05	5.15	3.68	4.00	1.37
	Total	75	70.14	4.48	6.77	3.31	8.61	5.54	3.49	3.50	1.40
First grade	GP	47	84.17	4.31	7.81	2.95	13.55	5.77	4.00	4.00	1.04
	FB	43	84.86	5.51	8.74	3.37	11.98	5.99	3.51	3.00	1.06
	Total	90	84.50	4.91	9.30	3.19	12.80	5.90	3.77	4.00	1.07
Second grade	GP	44	96.16	5.69	11.73	3.30	16.61	4.58	3.77	4.00	1.26
	FB	43	96.98	4.86	10.57	3.24	16.00	4.60	3.67	4.00	1.46
	Total	87	96.56	5.28	11.14	3.30	16.31	4.57	3.72	4.00	1.35
Total		252	84.45	11.65	9.19	3.69	12.78	6.17	3.67	4.00	1.27

^{a}in months, ^{b}raw score, ^{c}ordered category A-E (E=1,D=2,C=3,B=2,A=1) for the posttest. The correlation of the pretest and posttest proportion correct across individuals was $r = .65, p < .001$.

5.3.3 IRT *analyses per testing session*

The independent Rasch (1 PL) model parameters were estimated for the pretest and posttest using the Marginal Maximum Likelihood (MML) estimation procedure ($\theta \sim N(0, 1)$) from the `ltm` package for R (Rizopoulos, 2006). A parametric Bootstrap goodness-of-fit test using the Pearson's χ^2 statistic was used to investigate model fit, using the same `ltm` package. Based on 50 generated datasets the Rasch model fit of the pretest and posttest are acceptable ($p = .18$ and $p = .08$ respectively). The correlation between the item difficulty parameters for the item isomorphs of the pretest and posttest was strong: $r = .95$.

5.3.4 *Explanatory* IRT *analyses*

Each of the hypotheses about the children's performance and change on the 20 test items of the pretest and posttest sessions were investigated using model comparison. We first started with a simple IRT model. Predictors were then added successively and the fit of the new model was compared to the previous one. Because the previous restrictive model was nested in the new one, a likelihood ratio (LR) test could be used to test the improvement in goodness of fit. Each of these models was estimated using the `lmer4` package for R (Bates & Maechler, 2010) as described by De Boeck, et al. (2011). Table 5.2 presents an overview of comparisons between the estimated models; these are discussed in detail below.

Null model

The initial reference model (M0a) is a simple IRT model with random intercepts for both persons and items (pretest and posttest) where the probability of a correct

Table 5.2 *Overview of the estimated* IRT *models.*

		Effects								
	Nested		Random	Random					LR-test [a]	
Model	Model	Fixed	over Persons	over Items	AIC	BIC	LL	# p	df	Λ
M0			Intercept	Intercept	8667	8689	4330	2		
M1a	M0	+ Session	"	"	7581	7610	3786	4	2	1088.10***
M1b	M1a		+ Session	"	7381	7424	3684	6	2	204.07***
M1c	M1b		"	+ Session	7366	7423	3675	8	2	19.03***
M2	M1b	+ NrTransformations	"	Intercept	7354	7405	3670	7	1	28.42***
M3	M2	+ Session * Condition	"	"	7348	7413	3665	9	2	10.36**
M4a	M3	+ Age	"	"	7227	7299	3604	12	1	122.86***
M4b	M3	+ WorkingMemory	"	"	7260	7339	3619	11	2	92.19***
M4c	M4b	+ WorkingMemory	"	"	7204	7291	3590	12	2	26.89***
M4d	M4c	+ Session * Age	"	"	7209	7317	3590	15	3	1.16
M4e	M4c	+ Session * WorkingMemory	"	"	7211	7326	3590	16	4	1 .21
M5	M4c	+ Session * Math	"	"	7153	7254	3562	14	2	55.68***

[a] The LR-test comprises a comparison between the model and the nested model. Note: The LR-test is not always applicable when comparing random effects as the estimate is too conservative (De Boeck et al., 2011), however given the small p-value this is not a problem in the current situation.

*** p < .001, ** p < .01, * p < .05

response of person p on item i is expressed as follows.

$$P(y_{pi} = 1|\theta_p, \beta_i) = \frac{exp(\theta_p - \beta_i)}{1 + exp(\theta_p - \beta_i)} \tag{5.1}$$
$$\text{where } \theta_p \sim N(0, \sigma^2_\theta) \text{ and } \beta_i \sim N(0, \sigma^2_\beta)$$

It is common practice in the psychological literature to consider persons a random variable, based on the assumption that the participant was randomly selected from the population ($\theta_p \sim N(0, \sigma^2_\theta)$. A similar argument can be applied to items when these are drawn from a population of possible items as it is common practice in statistical models to use a normal distribution for residuals (De Boeck, 2008). In the present test the items can be considered a random sample selected from a pool of items that test figural analogical reasoning ($\beta_i \sim N(0, \sigma^2_\beta)$, rather than a definitive representation, which is important in the explanatory context when including factors that account for item difficulty (e.g., Baayen, Davidson, & Bates, 2008; De Boeck, 2008). We also conducted the same analyses with fixed item effects and reached the same substantive conclusions.

Model of learning and change

Our first research question focused on the effect of repeated testing. The first addition we tested against the null model was the inclusion of a session parameter to model average change from pretest to posttest. This resulted in M1a, which, as can be seen in Table 5.2, led to a significant improvement in model fit thereby confirming hypothesis 1a. M1a results showed that a child with average ability improved from having a probability of .06 to .33 in correctly solving an item of average difficulty from pretest to posttest ($B = 2.06, SE = .07, p < .001$).

Model M1a assumes the effect of retesting to be equal for all children (Fischer, 1976). In order to allow for individual differences in improvement from pretest to posttest, we applied Embretson's Multidimensional Rasch Model for Learning and Change (MRMLC) by including random parameters that allow for the session effect to vary over persons (e.g., Embretson, 1991b; Von Davier et al., 2010). As with the Rasch model, here the chance that an item is solved correctly (P_{ip}) also depends on the difference between the examinee's latent ability (θ_p) and the item difficulty (β_i). Yet, the ability is built up through the testing occasions m up to k in a summation term, which indicates which abilities (θ_{pm}) must be included for person p on occasion k.

$$P(y_{ipk} = 1|\theta_{pk}, \beta_i) = \frac{exp(\sum_m^k \theta_{pm} - \beta_i)}{1 + exp(\sum_m^k \theta_{pm} - \beta_i)} \tag{5.2}$$

$$\text{where } \theta_{pm} \sim N(0, \sigma_\theta^2) \text{ and } \beta_i \sim N(0, \sigma_\beta^2)$$

The initial ability factor, θ_{p1}, refers to the first measurement occasion (i.e. pretest) and the so-called modifiabilities (θ_{pm} with $m > 1$) represent gains from the previous test occasions. In the present model $k = 2$ and the modifiability θ_{p2} refers to performance change from pretest to posttest.

Including random modifiabilities in model M1b led to further improvement in model fit evidenced by lower AIC and BIC values and a highly significant LR-test. We could therefore statistically infer that individual differences in change from pretest to posttest were present, supporting hypothesis 1b. The variation of the children's improvement from pretest to posttest was rather large, $\sigma_2 = 2.25$. The children's modifiability scores showed a moderate negative correlation with their ability scores ($r = -.53$) indicating that children with lower pretest scores tended to improve more (see Figure 5.2).

However, note that the item difficulties (β_i) in Equation 2 are considered constant over occasions. This indicates that measurement invariance (cf. Meredith, 1993;

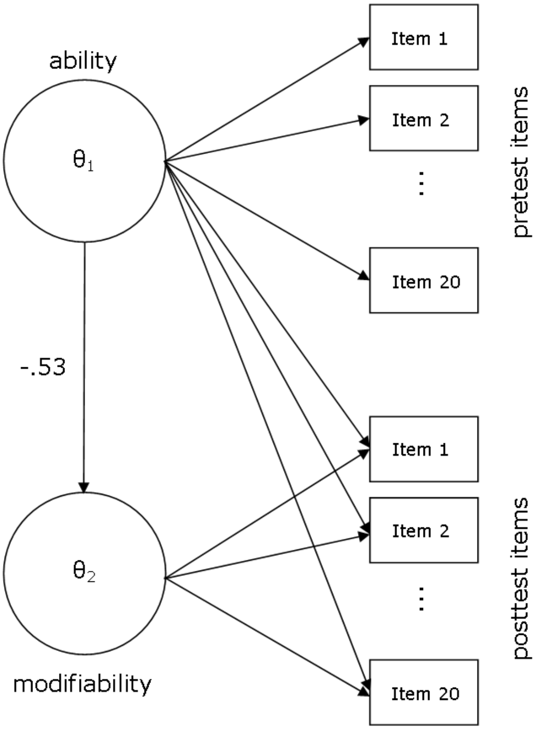

FIGURE 5.2 *Structural equation model of the relationship between ability and modifiability from* MRMLC *(Embretson, 1991a) applied in Model M1b.*

Millsap, 2010) is assumed within this model. In order to be sure that the effect of session was a global effect and not due to the items functioning differentially on the pretest and posttest (i.e. measurement invariance), we tested a model in which the session effect was allowed to vary over items. This model, M1c, improved model fit. However, the random item effects of the two sessions, $\beta_{pretest}$ and $\beta_{posttest}$, were highly correlated ($r = .97$). Hence we concluded that the session effect was global and we have therefore continued with M1b.

Modeling item difficulty

We tested whether our model could be improved by restricting the item difficulties to a linear combination of item variables (e.g., Janssen, Schepers, & Peres, 2004). As can be seen in Table 5.2 model M2, adding the number of transformations per item as a predictor improved model fit. The results show that for each additional transformation the children's chances of solving an item correctly decreases by .44 odds ($B = -.83, SE = .11, p < .001$).

Sources of individual differences in learning and change

Our model could be extended with more explanatory factors (De Boeck & Wilson, 2004; Hickendorff, Heiser, Van Putten, & Verhelst, 2008) by including other predictor variables and evaluating their effects on the latent scale. M2 includes person predictors for ability and modifiability (i.e. performance change from pretest to posttest) from MRMLC as well as a predictor of item difficulty. In the following analyses other person predictors (i.e. training-type, age-group, WMC, school performance) are included in order to explain the children's performance and change on the figural analogies scale. Person predictors are denoted as $Z_{pj}(j = 1, ..., J)$ and have regression parameters ζ_j. The item predictor (i.e. number of transformations) is denoted as $X_i(k = 1)$ and has the regression parameter δ. These explanatory parts are entered into the null model (see formula 1) as follows, with indices *i* for items, *p* for persons, *j* for the person covariate used as a predictor variable and *k* for the item covariate used a predictor variable.

$$P(y_{pi} = 1|Z_{p1} \ldots Z_{pJ}, \beta_i) = \frac{exp(\sum_{j=1}^{J} \zeta_j Z_{pj} + \epsilon_p + \delta X_{ik} + \epsilon_i)}{1 + exp(\sum_{j=1}^{J} \zeta_j Z_{pj} + \epsilon_p + \delta X_{ik} + \epsilon_i)} \tag{5.3}$$

Note that the person-by-session and item specific error parameters, ϵ_p and ϵ_i respectively, are assumed to stem from the normal distribution, i.e. $\epsilon_p \sim N(0, \sigma^2_{\epsilon p})$ and $\epsilon_i \sim N(0, \sigma^2_{\epsilon i})$. The results of which are presented in the following sections.

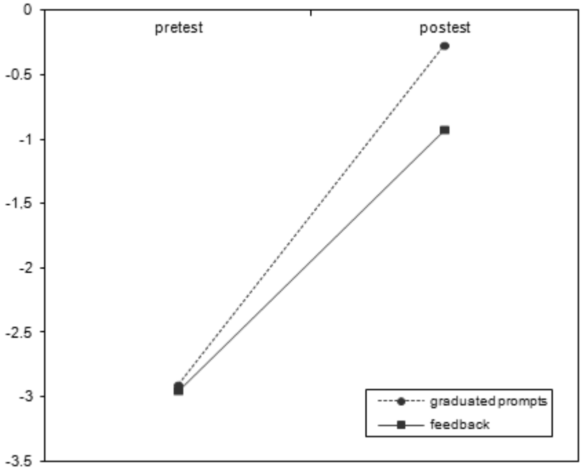

FIGURE 5.3 *Plot of person logits on an average item (four transformations) for both training conditions from pretest to posttest (M2b).*

Training effects. Our second research question was whether training with graduated prompts led to greater improvement on the analogical reasoning scale than training with feedback only and whether this was moderated by initial ability. To test this training-type x session was added as a predictor (M3). As a consequence, model fit improved, indicating that differences in performance between the two conditions were present. The main effect of session was $B = 2.64, SE = .17, p < .001$ (reference=pretest). The modifiability x training-type interaction effect was $B = -.61, SE = .24, p = .011$ (reference=graduated prompts). Simple contrasts showed that the effect was $B = 2.66, SE = .17, p < .001$ for the graduated prompts condition and $B = 2.00, SE = .17, p < .001$ for the feedback condition. A main effect for condition was not present ($B = -.05, SE = .32, p = .883$). As can be seen in Figure 5.3, children trained with graduated prompts (GP) showed greater gains than those in the feedback (FB) condition. The odds of solving an item with an average difficulty correctly increased by a factor of .52 for a child with an average ability in the graduated prompts condition, whereas this was .27 for an average ability child in the feedback condition. Here we also found that the children's modifiability scores in both conditions showed a moderate negative correlation with their pretest scores ($r_{GP} = -.51$ and $r_{FB} = -.46$), indicating that children with lower pretest scores tended to improve more, confirming hypothesis 3b.

Effects of age and working memory. The third research question aimed to investigate whether age or working memory or a combination best moderates children's performance on the dynamic test in question. We tested two models in which age-group (M4a) and working memory (M4b), were added as separate predictors. Of these two, M4a had the better fit (see AIC/BIC values in Table 5.2). Next we investigated whether WMC was an additional predictor by adding this to M4a; this improved model fit. In M4c both age-group and WMC had significant main effects. A positive relation between age and test performance was found ($B = .92, SE = .12, p < .001$), indicating that older children tended to have higher scores. Furthermore, verbal and visuo-spatial WM were significant predictors of analogy solving: $B = .36, SE = .12, p = .004$ and $B = .37, SE = .12, p = .002$ respectively. A positive relation between WM scores and performance on the figural analogies was present; the greater the WM scores the higher the performance estimates.

We tested whether age-group or WMC could explain individual differences in performance change from pretest to posttest by evaluating the interaction of the modifiability with each of these variables. Age did not interact with modifiability in a significant way: $B = -.05, SE = .17, p = .75$. The interaction effect of WMC and modifiability was also not significant: $B = .07, SE = .14, p = .61$ and $B = -.10, SE = .14, p = .50$ for verbal or visuospatial WMC respectively. In both cases model fit did not improve with explanatory factors for modifiability (see Table 5.3 models M4d and M4e). This means that the children's degree of improvement from pretest to posttest was not related to their age or WM scores.

Modifiability and math achievement. Finally we investigated whether modifiability was related to prior school performance in the form of achievement rating on a national standardized math assessment. Both the main effect of prior math achievement (Z-scores) and its interaction with modifiability was significant:$B = .45, SE = .13, p = .01$ and $B = .25, SE = .12, p = .04$ respectively (see Table 5.3 model M5). This means that the odds of solving an average item correctly by an average ability child increased by 1.57 odds per achievement level (1-5) increase if we assume that achievement is a continuous variable. We could conclude that the children's degree of improvement from pretest to posttest was significantly related to math achievement scores.

Final model

The best fitting IRT Rasch-scaled model (M5) shows significant fixed effects for session, WMC, age-group and prior math achievement as well as a significant interaction between session and training-type and also between session and math achievement (see Table 5.3). Random intercepts were present for persons per session ($SD_{ability} = 1.77, SD_{modifiability} = 1.44; r = -.72$) and items ($SD = .79$).

Table 5.3 *Estimates of fixed effects in model M5b.*

	B	SE	*p*
Intercept	40	50	.431
Session (reference = graduated prompts)	2.60	.17	<.001
Condition (reference = pretest)	.05	.26	.853
Session x Condition	-.60	.24	.011
Nr Transformations	-.83	.11	< .001
Age	.92	.11	<.001
Verbal WMC	.18	.11	.073
Visuospatial WMC	.28	.11	.009
Math	.45	.13	<.001
Session x Math	.25	.12	.038

In sum, these results indicate the following. Children generally improved from pretest to posttest, and individual differences in modifiability were present, confirming hypothesis 1. In accordance with hypothesis 2, the graduated prompts training led to a larger improvement in analogy solving compared to the feedback condition, although children with lower ability generally had greater modifiabilities. Investigations concerning research question 3 showed that age is related to performance, where older children solved the analogies better than younger children. Performance was also related to verbal and visuospatial WMC, where children with greater WMC obtained higher scores. Modifiability however was not related to age or WMC. Math achievement was related to analogy solving ability and modifiability, where children with higher math scores also performed better on the pretest and improved more from pretest to posttest.

5.4 Discussion

In the present study, the aim was to investigate children's differences in learning during a dynamic test of figural analogical reasoning using explanatory IRT models (De Boeck & Wilson, 2004). As with previous research on children's analogy solving progression, performance generally improved over repeated testing occasions, but the degree of improvement varied greatly (e.g., Freund & Holling, 2011b; Mackey et al., 2010; Siegler & Svetina, 2002; Tunteler & Resing, 2007c, 2007b). The large individual differences in learning and change after a short intervention coincides with findings in other cognitive tasks such as visuospatial reasoning (Embretson, 1987), series completion (Resing, Xenidou-Dervou, et al., 2012) and numerical estimation (Siegler, 2006; Luwel et al., 2010). The type of intervention, i.e. practice or training-type, appears to be one of the factors that influences these individual differences. We found that training with graduated prompts techniques, which includes metacognitive and strategy-based instructions, was significantly more effective in improving the children's analogy solving than feedback-training. This corresponds with the findings of Luwel et al. (2010) where strategy-feedback led to greater improvement in children's numerosity judgment than outcome-feedback. In the case of Luwel et al. (2010) especially children with lower intelligence test scores improved more with strategy-feedback. Also, Jaeggi et al. (2008) found that low ability children tended to improve more on figural matrices after training on a working memory task. Similarly, we found that children with lower pretest scores generally improved more, which given the moderate difficulty of the test items and the use of IRT estimations could not be due to ceiling effects. We therefore concur with the findings of Swanson and Lussier (2001) who concluded that children with initially lower cognitive ability scores tend to improve more during short dynamic testing training-phases. This indicates that children with untapped potential for learning are more often present in groups of low functioning children, but would perhaps be overlooked if they were judged based on a conventional, static reasoning test. Identifying these low functioning children with high potential for learning would be a necessary first step in helping them more fully realize their cognitive potential at school.

We investigated whether age or working memory affected performance on the dynamic test and found that older children generally performed better than the younger children (e.g., Siegler & Svetina, 2002; Sternberg & Rifkin, 1979; Tunteler & Resing, 2010), but that this was partly confounded by their working memory capacity. The combination of age and working memory capacity (WMC) was the best predictor of analogical reasoning pretest scores. Research has linked children's performance on fluid reasoning tasks, such as figural matrices, to their memory span and working memory capacity (e.g., Hornung et al., 2011; Kail, 2007; Tillman et al., 2008); therefore the contribution of WMC was not surprising. Yet as with two previous dynamic testing studies WMC was related to initial ability but unrelated to children's differences in improvement from pretest to posttest (Resing, Xenidou-Dervou, et al., 2012; Stevenson et al., submitted 2011a). Training fluid reasoning may improve working memory (Mackey et al., 2010). Therefore, we hypothesize that the short but adaptive training forms provided in these dynamic tests offers practice or problem solving strategies that aides in the more efficient use of the available working memory capacity. Including WMC measures both before and after training may help determine whether working memory efficiency is affected by the graduated prompts intervention, which is a task for future research.

Another related variable we investigated was whether school performance in math coincided with analogy solving and improvement during dynamic testing. Both initial ability and change scores were significantly related to math achievement. Previous research has demonstrated the relationship between fluid reasoning and math achievement (Primi et al., 2010; Taub et al., 2008). Support for the relationship between performance change and math achievement can be found in studies on the additional predictive value of dynamic outcomes for school performance (Beckmann, 2006; Caffrey et al., 2008; L. S. Fuchs et al., 2008). Perhaps dynamic testing outcomes are particularly suited in explaining individual differences in learning and achievement, i.e. developing expertise, over time (e.g., Swanson, 2011a; Stevenson et al., submitted 2012b). This should be addressed in conjunction with the role of working memory (e.g., De Smedt et al., 2009; Swanson, 2011b) in subsequent studies.

5.4.1 *Methodological implications*

In this paper we have argued that IRT is a helpful tool in the measurement of learning and change as it can provide gain scores without the statistical pitfalls classical test theory analyses suffer from (e.g., De Bock, 1976; Embretson, 1991b). In this study we extended Embretson's (1991b) Multidimensional Rasch Model for Learning and Change with an explanatory component and demonstrated the usefulness of De Boeck & Wilson's (2004) explanatory IRT approach in a dynamic testing context. This can easily be applied to other educational or developmental psychology research. This method holds great promise for dynamic testing and other intervention-based research, not only in reliably measuring differences in individuals' ability to learn, but also in explaining the sources of these differences.

The explanatory IRT context enables not only investigation of sources of variance in persons but also in sources of item difficulty (De Boeck & Wilson, 2004). We demonstrated that including the number of transformations in an analogy item improves the prediction of performance on an item. By including random item effects we treated the test items as being randomly drawn from a population of figural analogy matrices and also accounted for the item properties not perfectly explaining item difficulty (De Boeck, 2008). Modeling with fixed item effects led to the same substantive conclusions. However, including random item effects had the advantage of a more parsimonious model. In the present instrument design it was not possible to test the difficulty of each transformation separately as the transformation types were not counter-balanced per difficulty level. However, differences are expected, such as color being easier for children to identify and apply than orientation (e.g., Rijmen & De Boeck, 2001; Siegler & Svetina, 2002; Stevenson et al., 2011), and should be investigated in future studies.

We assessed whether measurement invariance was present as the psychometric properties of the test scores should not change per testing occasion when analyzing learning and change (Millsap, 2010). We found that the item parameters of the pretest and posttest were sufficiently related to directly compare the testing sessions in one IRT model. However, this is not always the case (e.g., Freund & Holling, 2011a; Lievens, Reeve, & Heggestad, 2007) and dynamic testing and intervention studies should address this issue when evaluating performance change over time.

5.4.2 Conclusion

Dynamic testing can be said to provide insight into an individual's learning potential through measures such as performance change from pretest to posttest (e.g., Embretson, 1987, 1992; Resing, 1997; Stevenson et al., submitted 2011a), instructional needs and strategy progression (e.g., Bosma & Resing, 2012; Bosma et al., submitted; Resing, 1997; Resing & Elliott, 2011; Stevenson et al., under review) and transfer (e.g., Campione et al., 1985; Stevenson et al., submitted 2011a). In the present study we analyzed sources of children's differences in performance change from pretest to posttest on a dynamic test of analogical reasoning. We found large variations in children's performance change and these were only partly related to initial ability, unrelated to WMC, but coincided with math achievement. This may indicate that performance change, measured with item response models, is an important construct in the assessment of learning and cognitive potential. Further research should focus on the relevance of dynamic testing outcomes in psychoeducational assessment – whether this indeed helps us measure and understand individual differences in cognitive capacity and potential.