Puzzling with potential : dynamic testing of analogical reasoning in children

Stevenson, C.E.

Puzzling with Potential

Dynamic testing of

analogical reasoning in children

Claire Elisabeth Stevenson

Stevenson, Claire Elisabeth
Puzzling with Potential: Dynamic testing of analogical reasoning in children.

Printed by Iskamp drukkerij, Amsterdam

ISBN 978-94-6191-395-1

Puzzling with Potential

Dynamic testing of

analogical reasoning in children

PROEFSCHRIFT

ter verkrijging van
de graad van Doctor aan de Universiteit Leiden,
op gezag van de Rector Magnificus Prof. mr. P. F. van der Heijden,
volgens besluit van het College voor Promoties
te verdedigen op 13 september 2012
klokke 13:45 uur

door
Claire Elisabeth Stevenson
geboren te Baton Rouge, Verenigde Staten

PROMOTIECOMMISSIE

Promotoren:

Prof. dr. W. C. M. Resing

Prof. dr. W. J. Heiser

Overige leden:

Prof. dr. J. G. Elliott	(University of Durham, UK)
Prof. dr. R. R. Meijer	(University of Groningen)
Prof. dr. K. H. Wiedl	(University of Osnabrück, Germany)

Contents

List of Figures

List of Tables

Acknowledgments

This page is dedicated to thanking the people both near (Leiden University) and far (Amsterdam, Groningen, USA, Sweden, Indonesia, ...) who supported me on this endevour.

I was very fortunate that Professor Wilma Resing chose to supervise me on this research project. Her knowledge and enthusiasm for the (dynamic) testing of children's cognitive development has been an inspiration. Wilma, your confidence in my abilities and the accompanying freedom to conduct the (numerous yet unreported) studies during my phD appointment have been greatly appreciated! Professor Willem Heiser, you have been so kind and a great support in helping me find my way in the academia. Thank you for initiating a collaboration with Professor Paul de Boeck and dr. Marian Hickendorff, who have become my IRT mentors. Paul, I especially want to thank you for teaching me the ins and outs of the explanatory IRT paradigm and I look forward to further collaboration in the future. Marian, your criticial comments regarding both data analysis and writing have been so helpful and I really enjoy(ed) our walks and talks.

I would like to thank my paranymphs, Evelien and Sheida, you two have been incredibly supportive throughout the years. Also, my collegues, both old (Leiden University) - especially Kirsten and Tirza from the Dynamic Testing Lab - as well as new (VU), thank you for being there. Dear friends and family, thanks so much for your love, support and letting me test you and your children's analogical reasoning skills! Mom, you are the world's best listener and editor – this dissertation is a tribute to you. Then last but not least, dear David, thank you for all of the post-its!

CHAPTER 1

General Introduction: Gathering the puzzle pieces

Nowadays, assessment is an integral part of our educational system. Although assessment procedures are ubiquitous in children's schooling, from preschool to university, measuring potential for learning remains a puzzle. Children are tested often during elementary school, informally with classroom quizzes or more formally by means of nationally normed achievement measures of school subjects such as math, reading and science. When educators suspect problems with regard to a child's learning or progression, they may inform a school guidance service or school psychologist who can administer an intelligence test and other psychodiagnostic instruments that measure cognitive abilities. School psychologists use these tests because the scores have considerable predictive value for school achievement and can be used as input for diagnoses of learning difficulties (Resing, 1997). However, some researchers from the fields of psychology and education have noted that this form of testing may not be the best manner to assess how well a child can learn. These conventional tests can underestimate cognitive ability – especially in disadvantaged groups such as ethnic minorities or learning disabled. In addition, they do not provide enough information that educators can use to create programs to remedy a child's learning problems (Grigorenko, 2009a; Haywood & Lidz, 2007). The concern is that conventional tests measure only what a child has learned up until the time of testing, but not his / her *potential for learning*, which is of particular interest when making decisions that impact a child's future education (Resing, 2000; Elliott, 2003). In order to address these shortcomings, some researchers have turned to dynamic testing, which examines an individual's ability to learn (Grigorenko & Sternberg, 1998). This introductory chapter describes the steps taken to develop a dynamic test of analogical reasoning for elementary school children while introducing the concepts and research questions addressed in this thesis.

Dynamic testing can best be described by contrasting it with a testing situation such as a scholastic achievement test taken for admittance to secondary school or

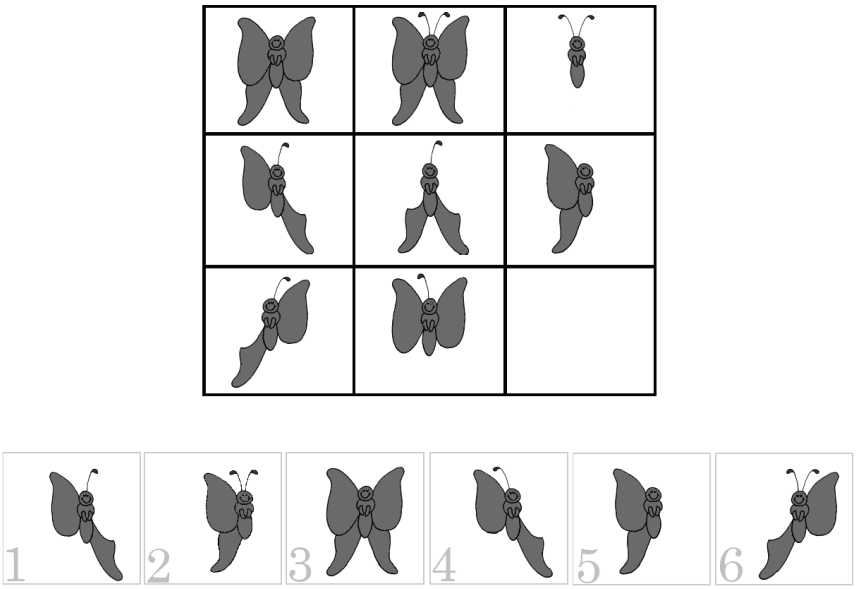

FIGURE 1.1 *An example item adapted from a cognitive ability test.*

university, or perhaps a cognitive ability test administered during the selection process for a job. An item similar to that of a cognitive ability test is shown in Figure 1.1. In a traditional testing situation, one may receive a short instruction: "Solve the following problems by selecting the option below that belongs in the empty box." An example problem may be provided; then the test-taker is asked to proceed solving a number of such problems without receiving further help or feedback.

The main difference between this typical "static" testing situation and that of dynamic testing is that training is incorporated into the dynamic assessment process. The examiner may inform the test-taker of the correct solution, in this case option 2, or more specific instruction may be provided. For example, if one tries to solve the reasoning problem in Figure 1.1, she or he may notice that an underlying rule

determines how the objects change horizontally and vertically. In this case, if a wing appears in two subsequent pictures then it is not present in the third. The same rule applies for the antennas, but the butterfly's body is present in each picture. This instruction may help the test-taker solve subsequent problems such as the one presented in Figure 1.2. The ability to profit from training in solving analogies and other cognitive tasks varies greatly between different individuals. The idea behind dynamic testing is that an individual's ability to profit from instruction provides an indication of one's cognitive potential (Elliott, Grigorenko, & Resing, 2010). Thus the amount of instruction someone requires to learn to solve the problems may be considered a way to measure this. For example, if an individual is unable to solve Figure 1.2 with only one previous example, perhaps more training would be useful. However, one should fear not if these items are still difficult; assessing the readers' learning potential is not the goal of this thesis.

Dynamic testing seems to provide useful information for educators about the learning potential of their students. For example, dynamic test results may be a useful addition to conventional tests in the prediction of scholastic achievement (e.g., Caffrey, Fuchs, & Fuchs, 2008) and can give more process-oriented diagnostic information that may help educator's intervene and improve an individual's performance at school (e.g., Bosma & Resing, 2012; Jeltova et al., 2011). Dynamic testing has shown much potential as an additional form of psycho-educational assessment; however, some hurdles prevent its wide-spread use (Sternberg & Grigorenko, 2002). One practical problem is that the dynamic testing process is more time consuming than traditional assessment methods due to extensive interventions incorporated into the assessment process. For this reason the dynamic test devised for this thesis has a relatively short training phase. A second problem is that the psychometric quality of dynamic tests is generally considered unclear or even poor. The main reason is that measuring potential ability is not simple – just

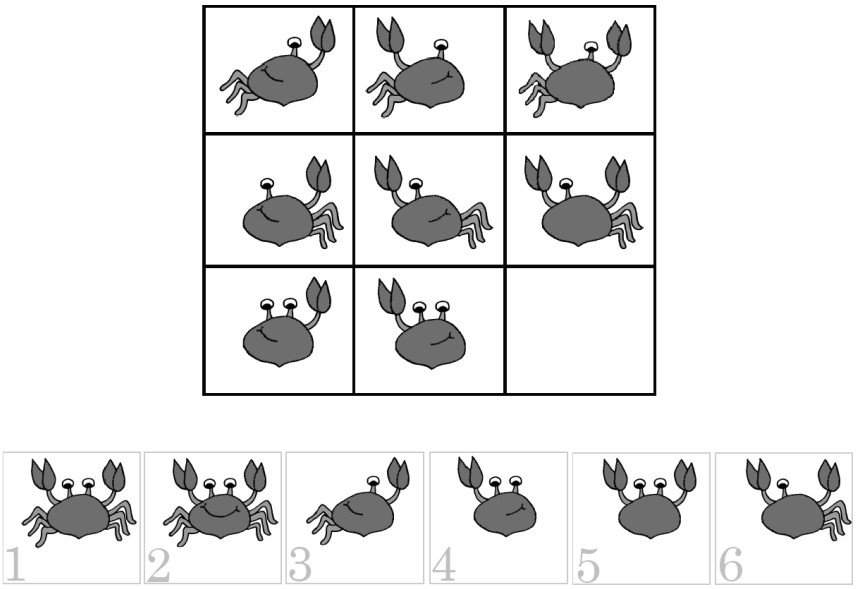

FIGURE 1.2 *A second example of a cognitive ability test item. The solution is option 5.*

as it may be difficult to predict a five-year-old's potential height from their height in the first week of kindergarten. Therefore, another goal of this research was to pay heed to rigorous psychometric standards, but still be able to provide valuable information unique to dynamic testing about an individual's learning process and cognitive potential.

Dynamic tests appear in various forms but have in common that some form of training is provided. In the dynamic test designed in this project a pretest-training-posttest format was used. Here the training method falls under graduated prompting techniques (e.g., Resing & Elliott, 2011). This form of training appears to be a good way to derive a picture of an individual's learning potential by looking at how well the child profits from instruction. It was first described in the context

of dynamic testing by Campione, Brown and colleagues (e.g., Campione & Brown, 1987; Ferrara, Brown, & Campione, 1986) and refers to the provision of increasingly specific instructions that aid the child in solving the problems. The prompts begin with a general instruction and then focus on improving metacognitive skills such as aiding planning or focusing attention on the task. If the child doesn't provide the correct answer an additional prompt is provided that explains problem solving steps – these are referred to as cognitive prompts. If this type of instruction is not enough, then the trainer guides the child to the correct solution in the form of scaffolds. The idea is that by using a standardized protocol and providing the smallest amount of instruction before each problem solving attempt, an individual's need for instruction to solve the problems can be gauged. The number of prompts required during training provides a measure of one's "Zone of Proximal Development" (ZPD, Vygostsky, 1978), i.e. the difference between what one is already capable of and that which can be accomplished with help of an instructor (Brown & French, 1979).

The dynamic test designed for this project, AnimaLogica, utilizes graduated prompting and is based upon the Learning Potential of Inductive Reasoning (LIR, Resing, 1990, 1993). Resing extended the graduated prompts approach in her test by using not only the number of prompts, but also the types of prompts that helped an individual most. The type of prompts, metacognitive, cognitive or scaffolding, provide an indication of which type of instruction a child best benefits from (Resing, 2000). A child's performance on the posttest then provided insight into his/her potential performance level. In addition, solution strategies were taken into account, which provide further information on *how* an individual's learns (Resing, Tunteler, De Jong, & Bosma, 2009). The idea with graduated prompting is that the training only temporarily leads to an improvement in performance. Still it provides measures and information on what a child's learning potential is and how this can be achieved, for example, in providing information to help educators

construct a treatment plan for a particular child (Bosma, Stevenson, & Resing, submitted).

The test developed in this project aims to continue a trend of providing insight into an individual's ability to profit from instruction by utilizing graduated prompting techniques. As with the LIR, inductive reasoning skills are assessed as these are considered central to intelligence (Carpenter, Just, & Shell, 1990) and essential for school learning (Goswami, 1992). In the present dynamic test, figural matrix analogies (see Figure 1.3) were used rather than the verbal analogies or visual exclusion task from the LIR. First of all, figural analogies are well suited for both computerized assessment and training, which would allow for more efficient test administration. A second reason was that figural analogies can be systematically constructed with a predictable difficulty level using rule-based item construction, which is helpful in selecting items that aren't too difficult or too easy for the intended audience (Primi, 2001). Thirdly, figural analogies are considered suitable for culture-fair testing (Cattell, 1979). Given the increasingly diverse cultural backgrounds of school children in the Netherlands, it was important to construct analogies with pictures of familiar objects that were expected to be less biased than for example verbal analogies.

The figural matrix analogies used in the dynamic test are a classical form of analogies (A:B::C:?) that are often used in psycho-educational assessment (see Figure 1.3). An example of an intelligence test that has been used throughout the world that comprises figural analogies is the Raven Standard Progressive Matrices (Raven, 1936). The ability to solve these types of problems develops with great variability throughout childhood (Leech, Mareschal, & Cooper, 2008). A child may start to learn to solve such problems only to make a number of errors on the next occasion and then gradually improve in further encounters. Improvement in solving analogies can take place spontaneously with practice (Tunteler & Resing, 2002).

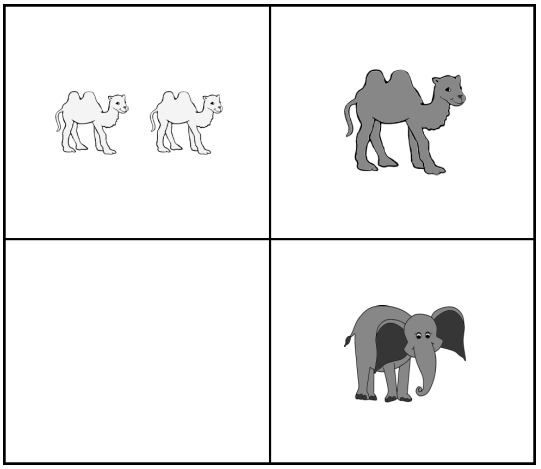

Figure 1.3 *An example figural analogy matrix item.*

Further learning effects can be found when individuals are provided with feedback on their solutions – i.e. told if their solution is correct or incorrect (Cheshire, Ball, & Lewis, 2005). Asking children to explain why they chose a particular solution provides additional benefit for their analogy solving progression (Siegler & Svetina, 2002). If a young child does not yet understand how to solve analogies, duplication errors are often made. A duplication error occurs when one of the analogy terms is copied – for example a duplicate solution of Figure 1.3 would be a copy of the red elephant in the lower right quadrant. When children improve in their analogy solving, they may make errors and provide only partially correct solutions. In Figure 1.3, a partial analogical solution would be two small red elephants – where the solution is almost correct except for one aspect such as color. As children get older, their correct analogical solutions increase, although the amount of change that takes place also differs greatly between children (e.g., Tunteler, Pronk, & Resing, 2008). A pilot study of this thesis (Stevenson, Resing, & Froma, 2009) addressed

the positive role of self-explanation, in addition to feedback, in children's strategy progression on the figural analogies.

Another pilot study (Stevenson, Touw, & Resing, 2011) investigated whether the figural analogies items were appropriate for the computer. It compared young children's solutions and strategy progression on computer versus paper figural analogies puzzles. The computer version had a clear advantage with regard to time investment– supporting our first reason for choosing figural analogies. Also, the difficulty level could be predicted as expected by the number of underlying rules, which helped us develop items for older children as well. However, another aspect of choosing appropriate items for dynamic testing was whether or not to use multiple-choice items. Multiple-choice items are generally easier to solve, but they do not provide a direct view of the strategies a child used to solve the item. This limitation makes it difficult to achieve one of dynamic testing's aims, which is to analyze an individual child's learning process – perhaps in order to diagnose errors in their reasoning. Therefore in Chapter 2 we addressed the role of item format in dynamic testing by comparing 5-6 year old children's performance during training on items with multiple-choice selection versus constructed-response, a type of open format question.

The suitability of the items and test for culturally diverse populations is addressed in Chapter 3. Dynamic testing is considered a promising method for multicultural assessment (Grigorenko, 2009b). On conventional tests, individuals from the dominant culture generally have an advantage compared to their peers with other ethnic backgrounds, for example due to factors such as test-wiseness or non-native instruction language. The presence of these factors can lead to a misrepresentation of ethnic minority children's cognitive potential. The idea is that through repeated practice or training, cultural differences become less prominent and disadvantaged children are provided with more opportunity to reveal their

cognitive potential (Sternberg et al., 2002; Van de Vijver, 2008). In Chapter 3 we addressed the question of whether the developed dynamic test is appropriate for culturally diverse schools by comparing indigenous Dutch and ethnic minority children's performance.

Two other factors that may play a role in the dynamic testing of analogical reasoning are age and working memory. On the whole adults are more capable of solving analogies than children, and older children tend to perform better than younger children. A possible explanation may be that the efficiency of working memory improves with age. Working memory is the ability to hold and manipulate entities in memory (Swanson, 2008). Because working memory has been shown to be related to the ability to solve matrix analogies in both adults and children (Kail, 2007; Kyllonen & Christal, 1990), it seems important to investigate its role in the dynamic testing of children's figural analogy solving ability. This question is addressed from different perspectives in Chapters 4 and 5.

In Chapter 4, the role of working memory is investigated in the context of transfer effects. Transfer is the ability to spontaneously generalize a problem-solving approach taught in one context (such as during training) to a different but applicable situation (Detterman, 1993). For example, if a person was trained in solving items such as those in Figure 1.3, then the items in Figures 1.1 & 1.2 could be considered a measure of near-transfer. However, transfer does not seem to occur easily as learning is context-bound and children do not often recognize that their acquired problem solving skills can be applied in new situations (Barnett & Ceci, 2002). Yet, transfer of skills to novel situations may provide additional insights into a child's potential for learning (Bosma & Resing, 2006). In Chapter 4 we investigated the extent to which the children were able to apply the reasoning skills learned during the dynamic testing of analogical reasoning to similar untrained tasks and explored the role of working memory herein.

In Chapter 5 we examined the roles of working memory, age and initial ability in the dynamic testing of analogical reasoning in more depth. In this chapter the main question was whether these factors interact with training type in explaining children's change in performance from pretest to posttest. However, in order to analyze these individual differences in performance change we first had to focus on a major obstacle in dynamic testing: how to obtain and interpret reliable measures of individual change when comparing performance before and after training (Sternberg & Grigorenko, 2002). The problem is that the change score is unreliable (Lord, 1963) if the change score is obtained by subtracting the percentage of correct solutions (or strategies or explanations) on the pretest from that on the posttest. Measuring change in this manner has received much criticism by psychometricians (e.g., Embretson, 1991b). Item response theory – a form of statistical modeling often employed in test design and educational measurement – provides a different way to estimate the scores of a child's performance on a dynamic test and does not suffer from the problem of reliability (Embretson & Reise, 2000). The analyses included in each of the studies reported in this thesis include item response theory analyses. However, Chapter 5 examined the measurement of performance change from pretest to posttest in greater detail. Here we used item response models not only to estimate the children's progression from pretest to posttest but also to explain which factors such as age, working memory, type of training or school performance were related to the differences between the children's performance change.

This thesis also addressed the value of dynamic testing outcomes in providing information relevant to children's learning at school. One way to demonstrate the value of a dynamic test is to investigate how well the results predict school performance (Beckmann, 2006). Conventional test results are generally considered good predictors of academic achievement but such prediction is found to be

less accurate in children (Sternberg, Grigorenko, & Bundy, 2001) even though performance on conventional tests with figural matrices is always somewhat related to school performance (e.g., Balboni, Naglieri, & Cubelli, 2010), as analogical reasoning is fundamental to school learning (Goswami, 1992). However, there is some evidence that dynamic tests may be of additional or better predictive value with regard to school performance (Caffrey et al., 2008; Hessels, 2009; Swanson, 2011a). Yet, it remains unclear whether present capacity (as measured with conventional tests) or the dynamic testing outcomes such as performance change from pretest to posttest, transfer ability or the number and type of required prompts during training best predict school performance. Therefore, in Chapter 6 the ability to solve analogies during conventional testing and dynamic testing are compared as predictors of school performance.

Summary

In sum, dynamic testing aims to provide a measure of abilities that are not yet fully developed by focusing on potential for acquiring new knowledge across multiple testing occasions. Instruction is incorporated into the training sessions which are preceded by a pretest and followed by a posttest. The static pretest provides an indication of present ability and the dynamic posttest shows what an individual may be capable when provided with tailored instruction. The number and type of instruction required to solve the problems during the graduated prompts training provides further information on an individual's potential for learning. The ability to solve and explain new but similar transfer problems may provide additional information about an individual's potential for learning.

The main research question of this thesis was: *"Which factors influence a child's performance on a dynamic test of analogical reasoning?"*. Chapter 2 addressed the influence of item format – whether training during dynamic testing differs when

using multiple-choice or constructed-response items leads to differences in children's analogy solving with regard to strategy progression, self-explanation or change in performance from pretest to posttest. The information gained from this study led to the decision to use constructed-response items in further studies. In Chapter 3 dynamic testing of culturally diverse school children is investigated. Here we found that performance change did not differ between indigenous Dutch and ethnic minority children. We did find that working memory measures did not differ between these two groups, but was related to the children's ability to solve figural analogies. Therefore Chapter 4 further examined the role of working memory in explaining individual differences in training and transfer effects of the presented dynamic test of analogical reasoning. The results seemed to indicate that working memory only plays a role in describing initial performance but is neither a factor in how children progress from pretest to posttest or in their transfer task performance. However, given the small sample size more research was required in order to draw any conclusions. The study presented in Chapter 5 therefore included more children and from three different age groups and further analyzed what leads to these individual differences in performance and change on a dynamic test of figural analogies. Performance change was found to be related initial ability but not working memory or age, yet was associated with math achievement. This finding led to the conclusion that performance change may be an important construct for educational psychologists in assessing school children's potential to learn and led to the examination of the predictive value of this construct. The issue of predictive value was studied in Chapter 6, where static and dynamic measures were compared in the prediction of children's school achievement in reading and math. This thesis concludes in Chapter 7 with an overview of the results of each of the preceding chapters and discusses potential answers to the question of which factors play a role in children's performance and change on a dynamic test of analogical reasoning.

CHAPTER 2

Dynamic testing of analogical reasoning in 5-6 year olds: multiple-choice versus constructed-response training

This chapter is based on Stevenson, C. E., Resing, W. C. M. & Heiser, W. J. (accepted conditionally upon revision). Dynamic testing of analogical reasoning in 5-6 year olds: multiple-choice versus constructed-response training. *European Journal of Psychological Assessment.*

Abstract

Multiple-choice analogy items are often used in cognitive assessment. However, in dynamic testing, where the aim is to provide insight into potential for learning and the learning process, constructed-response items may be of benefit. This study investigated whether training with constructed-response (CR), or multiple-choice (MC) items leads to differences in the strategy progression and understanding of analogical reasoning in 5-6 year olds (n=111). A pretest-training-posttest control group design with randomized blocking was utilized, where two experimental groups were trained according to the graduated prompts method. Results show that both training conditions improved more during testing compared to untrained controls. Children in the CR condition required more aid during training and showed different strategy-use patterns compared to the MC group. However, the quality of solution explanations was significantly better for children in the CR condition. It appears that performance advantages of training with CR items are most apparent when active processing is required. In the future, we advise including items that stimulate active processing and allow for fine-grained analysis of strategy-use, such as CR or analogy construction in dynamic testing to further discern differences in children's analogical reasoning understanding.

Acknowledgments

We would like to thank Carina de Klerk for her assistance with organizing and conducting the data collection and coding.

2.1 Introduction

Dynamic testing, often contrasted with static tests such as traditional IQ assessment, aims to provide a measure of abilities that are not yet fully developed (e.g., Elliott et al., 2010; Sternberg & Grigorenko, 2002). Where static tests measure previously acquired knowledge at one point in time, dynamic tests focus on potential for acquiring new knowledge across one or multiple testing occasions. Dynamic testing procedures further differ from static testing in that feedback is provided by the examiner in order to facilitate learning during assessment. Dynamic tests often consist of a pretest-training-posttest design where structured feedback is provided during one or more training sessions. The effectiveness of various types of training and feedback has been demonstrated in a dynamic testing context (e.g., Day, Engelhardt, Maxwell, & Bolig, 1997; Lifshitz, Tzuriel, & Weiss, 2005; Resing et al., 2009). However, not only feedback type influences strategy-use, learning and transfer (e.g., Luwel, Foustana, Papadatos, & Verschaffel, 2010), but also problem format. For example, open-ended items are generally found more difficult to solve (Behuniak, Rogers, & Dirir, 1996; Currie & Chiramanee, 2010; In'nami & Kozumi, 2009), but provide more diagnostic information (Birenbaum & Tatsuoka, 1987; Birenbaum, Tatsuoka, & Gutvirtz, 1992; Currie & Chiramanee, 2010; Martinez, 1999) and problem construction rather than multiple-choice solution may lead to greater learning and transfer (Harpaz-Itay, Kaniel, & Ben-Amram, 2006; Martinez, 1999).

In the current experiment, the aim was to investigate the effects of problem format in a dynamic testing context on learning and strategy-use. It was examined whether training using figural analogy problems, in which the solution must be constructed, would lead to greater progression in performance than training with multiple-choice (MC) problems in a dynamic test of analogical reasoning.

Dynamic testing is often conducted with analogical reasoning tasks (Resing, 2000) as analogical reasoning is considered a core component of intelligence (Carpenter

et al., 1990) and essential to school learning (Goswami, 1992). The various training formats used in dynamic tests generally show that children improve their skills through instruction and that posttest scores provide a better indication of their potential ability (Fabio, 2005; Sternberg & Grigorenko, 2002). Furthermore, utilizing graduated prompting techniques enables the determination of the amount and type of instruction a child requires to perform at this potential level (e.g., Ferrara et al., 1986; Resing, 2000; Resing & Elliott, 2011). In the case of inductive reasoning tasks, graduated prompting has been shown more effective than practice with regard to both accuracy and strategy development (Bosma et al., submitted; Ferrara et al., 1986; Resing, 1993; Resing et al., 2009). Training young children's analogical reasoning decreases duplication errors, in which one of the analogy terms is copied, and partial and correct analogical solutions increase with self-explanation, feedback (e.g., Cheshire et al., 2005; Siegler & Svetina, 2002; Stevenson et al., 2009; Tunteler et al., 2008) and graduated prompting (Tunteler & Resing, 2010). Although much research has been conducted on the effects of training on analogical reasoning, few studies have investigated the influence of task format on analogy learning and strategy development.

In the context of dynamic testing, item formats are interesting for two reasons. First, constructed-response (CR) items have been found to provide diagnostic advantages in determining where a pupil goes wrong if the solution is incorrect (Birenbaum & Tatsuoka, 1987; Martinez, 1999). This diagnostic information is valuable for process-oriented aims of dynamic testing such as examining strategy-use and instructional needs (e.g., Resing et al., 2009; Resing & Elliott, 2011). In the case of analogies, strategies, such as duplication or partially correct, can be determined directly rather than inferred from the limited multiple-choice (MC) options. Furthermore, diagnosis of systematic errors such as continually disregarding a specific transformation, e.g. orientation, can be more accurate as

the errors are not limited to the possible MC answers. The second reason is that problem construction formats may lead children to develop deeper understanding than using multiple-choice items (Bernardo, 2001; Harpaz-Itay et al., 2006).

Harpaz-Itay et al. (2006) found that analogy construction training led to better performance on verbal, geometric and numerical analogy tasks than training with MC items. They argued that MC solution is largely based on recognition, whereas construction employs conceptual task analysis. Response construction may also have advantages and evoke more complex thinking as the answer cannot be constructed based on recognition or response elimination (Bridgeman, 1992; Martinez, 1999).

Solving analogies and matrices with MC items is related to number and type of available options (Vigneau, Caissie, & Bors, 2006). Young children often rely on perceptual matching and are strongly influenced by the presence of distractors (Richland, Morrison, & Holyoak, 2006; Thibaut, French, & Vezneva, 2008), which can lead to a misdiagnosis of their understanding (Birenbaum et al., 1992; Goswami, 1992). These pitfalls could be said to fall under the response elimination method, where responses are tested until the best fitting option is chosen as the solution. This method is often used by those with weaker analogical reasoning skills, whereas constructive matching, where the problem is solved before constructing or selecting the solution, is usually employed by more advanced reasoners (Bethel-Fox, Lohman, & Snow, 1984; Vakil, Lifshitz, Tzuriel, Weiss, & Arzuoan, 2010). Constructive matching seems a prerequisite to consistently solve CR items correctly and teaching this strategy without the presence of distractors may be beneficial to children.

In this study we investigated the effectiveness of two training item types on the dynamic testing of analogical reasoning skills: constructed-response (CR) versus multiple-choice (MC). Our first research question concerned whether the graduated prompts training led to greater learning of analogical reasoning in young children

than solving a control task. In accordance with the literature we expected (1a) all children would improve in figural analogical reasoning with time, yet (1b) the graduated prompts training would add to this effect (Ferrara et al., 1986; Resing, 1990; Resing et al., 2009; Tunteler & Resing, 2010), leading to greater improvement in both training conditions compared to the control group. Our second research question focused on the effects of item format on performance during training. We expected (2a) the CR items to be more difficult than MC items (Behuniak et al., 1996; Currie & Chiramanee, 2010; Martinez, 1999), but (2b) that training with the CR format would lead to better understanding – revealed by better explanations of the solution – compared to MC. Finally we investigated item effects on strategy progression, by comparing strategy-use patterns of the two training conditions. We expected (3) CR-trained children to utilize more advanced analogical reasoning strategies, i.e. fewer duplications and more partial and correct solutions, than the MC-group both during training and on posttest measures (Harpaz-Itay et al., 2006; Resing & Elliott, 2011; Tunteler et al., 2008).

2.2 Method

2.2.1 Participants

Participants were 111 children (54% girls; M=64, SD=7 months). All children were native Dutch speakers, from two elementary schools in the Netherlands - selected based upon their willingness to participate. Written informed consent was obtained from the parents.

2.2.2 Design

A pretest-training-posttest control-group design with randomized blocking was employed. Children were blocked into one of three conditions: (1) training with

MC items, (2) training with CR items and (3) a control group. Randomized blocking was based on visual exclusion scores (Bleichrodt, Drenth, Zaal, & Resing, 1987), classroom and gender. All children solved the 20 pretest items during the first session. In the following two sessions trained children received the graduated prompts training with either MC or CR items. The children were trained on 4 items per session with 8 items total – limiting the duration of each session to 20 minutes. The control group solved maze coloring tasks. During the last two sessions, posttests - parallel versions of the pretest, were administered. Sessions took place weekly in a quiet location at the child's school, except for the last session which took place two weeks after the first posttest.

Visual exclusion

The RAKIT subtest Visual exclusion (Bleichrodt et al., 1987) measures inductive reasoning ability. The children must induce a rule to determine which figure does not belong.

ANIMALOGICA*: test and training*

The visual analogies material was based on the items utilized by (Stevenson et al., 2009) consisting of colored (red, yellow or blue) animal figures, classically presented in 2x2 matrix format. Drawings of familiar animals occupied three squares and the lower right or left quadrant was empty. Transformations comprised the dimensions: (1) animal, (2) color, (3) size, (4) position, (5) orientation and (6) quantity.

For the MC-items, used during the pretest, posttests and MC-training, the solution could be selected from five systematically constructed alternatives: (1) correct answer, (2&3) partial answer: missing one transformation, (4) duplicate answer: a copy of the term above or next to the empty box and (5) other non-analogical answer: missing two or more transformations (see Figure 2.1). In

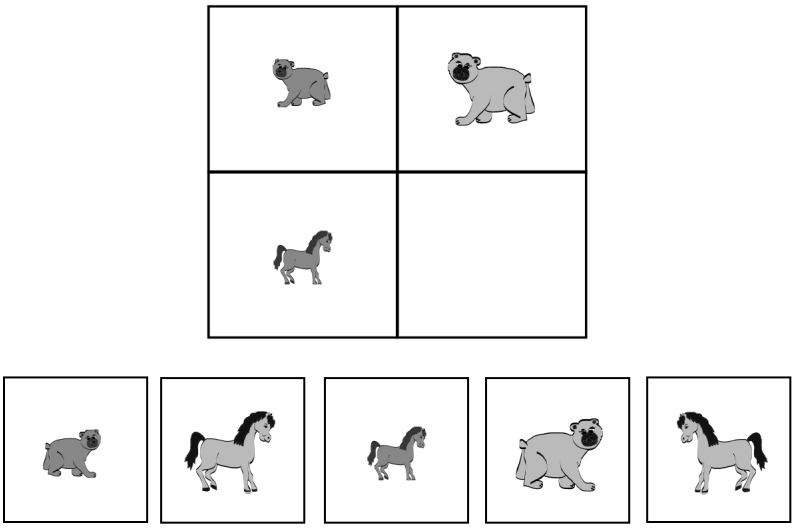

Figure 2.1 *Example MC item from* AnimaLogica *with options representing the strategies (from left to right) non-analogical, correct, duplicate, partial and partial respectively.*

the CR-training the solution was constructed from a number of animal cards representing the six transformations (see Figure 2.2); each animal was available in the three colors, two sizes (large, small) and printed on two sides, so by turning the card over the animal's orientation could be changed (looking left by default or turning over to look to the right). Quantity was specified by selecting one or more animal cards and position was selected by the placement in the empty square.

During training graduated prompting - a standardized, yet adaptive training procedure - was used (e.g., Bosma et al., submitted; Ferrara et al., 1986; Resing, 1993, 2000; Tunteler & Resing, 2010). Each item began with a general instruction. The examiner recorded the child's answer and if this was incorrect, a prompt was provided. If another mistake was made the next prompt, consisting of more specific instruction, was given. This stepwise approach begins with general,

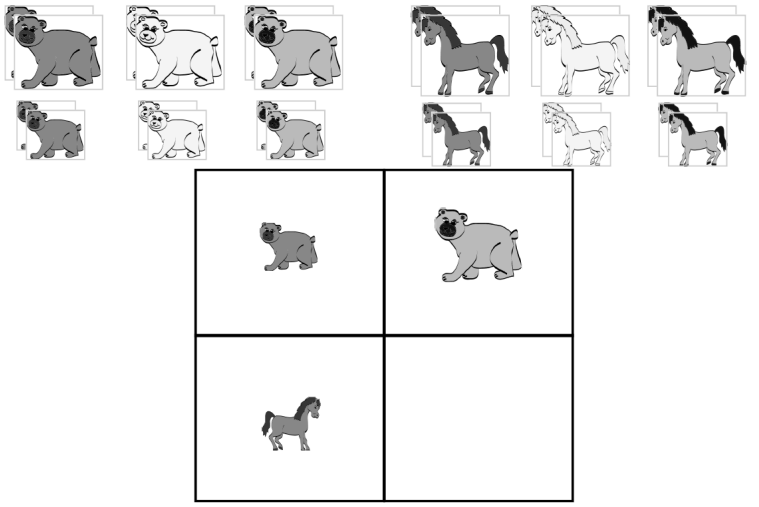

FIGURE 2.2 *Example CR item from* ANIMALOGICA.

metacognitive prompts, such as focusing attention, followed by cognitive hints, such as emphasizing the transformations in the item, and finally step-by-step scaffolds to solve the problem. Once the child answered correctly, he or she was asked to explain the correct solution. The trainer then provided an explanation of the solution – regardless of the correctness of the child's explanation. No further prompts were given and the next item was then administered.

2.2.3 Scoring

The children's analogy solutions were scored in two ways. First, scores based on correct/incorrect solutions were obtained using Rasch estimates from item response theory. Item response theory models were chosen as these seem to circumvent statistical problems (e.g., unreliability, scaling of change is not necessarily the same for persons with different pretest scores) encountered when using proportion correct

as the dependent variable in measuring performance change over time (Embretson & Reise, 2000). Rasch model scores are based on a person's ability as well as item difficulty. Rasch estimates were obtained for a joint logistic scale of pretest and posttests performance using Andersen's Rasch Model for repeated measurements (Andersen, 1985).

The second way the children's pretest and posttest solutions were categorized was into four strategies based on the literature (e.g., Cheshire et al., 2005; Siegler & Svetina, 2002; Tunteler & Resing, 2002; Tunteler et al., 2008) for analyzing strategy-use: (1) correct analogical solutions as correct answer selection or construction, (2) partial analogical were solutions missing one transformation, (3) duplicate non-analogical solutions were copies of the B or C term, and (4) other non-analogical solutions as answer choices missing more than one transformation (see Figure 2.1). A duplication error was always scored as category 3 – even if the duplicate was missing only one transformation.

Two measures were obtained from the graduated prompts training: (1) the number of prompts required per item and (2) quality of each child's explanations of the correct solutions. The children explanations of the correct solution of each training item were quantified by the number of correctly explained transformations (Stevenson et al., 2009). Furthermore, the categorization of the children's first solution to each training item was used in analyzes of strategy progression.

2.3 Results

2.3.1 Initial Group Comparisons

The children's initial level of inductive reasoning, measured with the visual exclusion task, did not differ between the three conditions according to an ANOVA ($F(2, 108) = .21, p = .814$). The average age per condition also did not differ ($F(2, 108) = .15, p =$

.860). Initial performance on the figural analogies was related to performance on the exclusion test ($r = .37, p < .001$) and age ($r = .41, p < .001$).

2.3.2 *Psychometric Properties*

The reliability of the pretest, $\alpha = .78$, is satisfactory. The reliabilities for the first posttest per condition were $\alpha_{MC} = .85$, $\alpha_{CR} = .90$ and $\alpha_{control} = .81$. The internal consistencies for the second posttest were $\alpha_{MC} = .88$, $\alpha_{CR} = .88$ and $\alpha_{control} = .85$. The reliabilities of the test on both sessions for each condition are considered good. The reliabilities of the training scale (8 items), calculated using the number of required prompts per item, are satisfactory: .83 and .78 for the MC and CR conditions respectively. The test-retest reliability for the control group three weeks after initial testing was, $r = .83$, $p < .001$ ($N = 39$), indicating good stability over time. The proportion correct of the pretest items ranged from .11 to .80 ($M = .31, SD = .42$); on the first and second posttest this was .23 to .91 ($M = .50, SD = .46$) and .23 to .95 ($M = .56, SD = .45$) respectively.

The independent Rasch (1 PL) model parameters were estimated for the pretest and posttests 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 fits of each test occasion using the same `ltm` package. The model fit of the first posttest was acceptable ($p = .36$). For the pretest and second posttest this was less satisfactory ($p = .04$ and $p = .04$). However, the item fit statistics for the items of both measurement moments were generally satisfactory ($p > .05$) and therefore the models were deemed acceptable. The correlation between the item difficulty parameters for the items of the pretest and first posttest was moderate, $r = .67$, and the correlation between the two posttests was strong, $r = .82$. We therefore considered the application of Andersen's Rasch model for repeated measurements

Table 2.1 *Basic statistics of Rasch ability estimates for figural analogies pretest and posttest.*

	Control (N=39)		MC Training (N=36)		CR Training (N=36)		Total (N=111)	
	M	SD	M	SD	M	SD	M	SD
pretest	-0.011	0.826	-0.252	0.892	0.159	0.974	-0.034	0.905
posttest 1	0.643	1.158	.929	1.461	1.140	1.290	0.897	1.309
posttest 2	0.954	1.301	1.255	1.600	1.440	1.412	1.209	1.440

appropriate. Fit statistics for the Andersen model estimated using the `lmer4` package for R (Bates & Maechler, 2010) were AIC = 6844, BIC = 7021, LL = −3396.96 with 26 parameters. The `ranef` function in the same package was used to extract the person Rasch-scaled estimates per testing occasion.

2.3.3 *General effect of training*

Our first research question concerned the effect of the graduated prompts training on young children's analogical reasoning. We expected (1a) all children's figural analogical reasoning to improve with time, but that (1b) trained children would show greater improvement. This was investigated using repeated measures (RM) ANOVA with Rasch-scaled ability estimates per session as dependent variable (see Table 2.1 for basic statistics), with session as within-subjects variable and condition as between-subjects variable. The analysis revealed a main effect for session (Wilks' $\lambda = .38, F(1, 108) = 177.12, p < .001, \eta_p^2 = .62$) showing that children, on average, progressed in figural analogy solving across sessions, confirming hypothesis 1a. The significant interaction effect for session x condition (Wilks' $\lambda = .92, F(2, 108) = 4.82, p = .010, \eta_p^2 = .08$) indicates that children in the conditions differed in degree of progression. Simple contrasts showed that both the CR and MC training-groups improved more than the control-group ($F(1, 73) = 4.31, p = .041, \eta_p^2 = .06$ and $F(1, 73) = 8.92, p = .004, \eta_p^2 = .11$ respectively), confirming hypothesis 1b.

2.3.4 *Comparison of Training Item Format: Prompting and Explanations*

Our second question pertained to the effect of training item format (MC or CR) on performance during the graduated prompts training. We hypothesized that (2a) CR items would be more difficult than MC items, but that at the same time (2b) CR-trained children would provide more advanced answer explanations.

To investigate the difficulty of the training items we analyzed the number of prompts required by the children. A RM ANOVA with number of prompts as the dependent variable, one within factor (item: 1 – 8), and one between factor (condition) was conducted. There was a main effect for item (Wilks' $\lambda = .37, F(7, 64) = 15.40, p < .001, \eta_p^2 = .63$) showing that children generally required fewer prompts during the course of training (see Figure 2.3, top). The significant item x condition interaction effect (Wilks' $\lambda = .65, F(7, 64) = 4.92, p < .001, \eta_p^2 = .35$) and significant between-subjects effect for condition ($F(1, 70) = 38.49, p < .001, \eta_p^2 = .36$) indicate that MC-trained children required fewer prompts than those trained with CR items, in accordance with hypothesis 2a.

Children's explanations of the correct solution were also analyzed using RM ANOVA with explanation quality as the dependent variable, one within factor (item: 1-8) and one between factor (condition). Again, there was a main effect for item (Wilks' $\lambda = .09, F(7, 64) = 89.12, p < .001, \eta_p^2 = .91$) showing that on the whole children used more advanced explanations during the training sessions (see Figure 2.3, bottom). The interaction effect for item x condition (Wilks' $\lambda = .73, F(7, 64) = 3.34, p = .004, \eta_p^2 = .27$) and significant between-subjects effect ($F(1, 70) = 12.25, p = .001, \eta_p^2 = .15$) show that children in the CR condition provided more advanced explanations compared to children in the MC condition, confirming hypothesis 2b.

2.3.5 Comparison of Training Item Format: Strategy-use patterns

Our third research question focused on the effect of training item format (MC or CR) on strategy-use patterns. Here we compare the strategies of the MC and CR training group across each of the dynamic test sessions. Children's solutions were categorized as correct, partially correct, a duplicate or other. We hypothesized that (3) training with CR items would lead to more advanced strategy-use than training with MC items.

As can be seen in the depiction of strategy progression in Figure 2.4, the children generally increase correct solutions from pretest to posttests and decrease incorrect strategies. Yet some differences between the two conditions seem apparent, especially during the training sessions. Changes in proportions of strategy-use across sessions were analyzed, as well as possible differences between MC and CR training conditions, with a MANOVA (2 conditions x 5 sessions) with repeated measures for session. The dependent variables were proportion strategy-use for the correct, partial and duplicate strategy. The other strategy was not included because of redundancy (i.e. the four strategies form a linear combination). There was a main effect for session (Wilks' $\lambda = .13, F(12, 59) = 34.32, p < .001, \eta_p^2 = .88$), which implies that strategy-use differed from session to session. A significant interaction effect for session x condition was present (Wilks' $\lambda = .58, F(12, 59) = 3.62, p = .001, \eta_p^2 = .42$) indicating that the two conditions differed in proportions of strategy-use across sessions, confirming hypothesis 3. MANOVAs per session with condition as factor and the 3 strategies as dependent variables were conducted in order to pinpoint when these differences occurred. A significant effect was found only for the first training session, Wilks' $\lambda = .71, F(3, 68) = 3.40, p = .001, \eta_p^2 = .29$. As can be seen in Figure 2.4, MC-trained children use more correct and duplication strategies, whereas partial strategies are most often applied during the CR training.

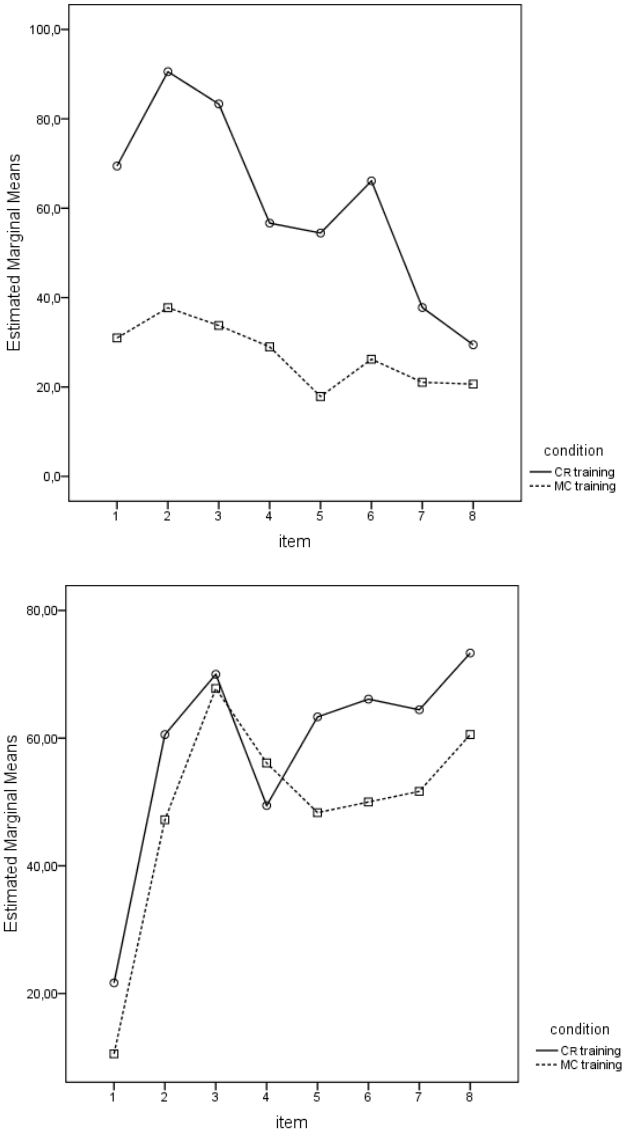

FIGURE 2.3 *Progression of required prompts (top) and explanations (bottom) per condition and across training items – both sessions are included.*

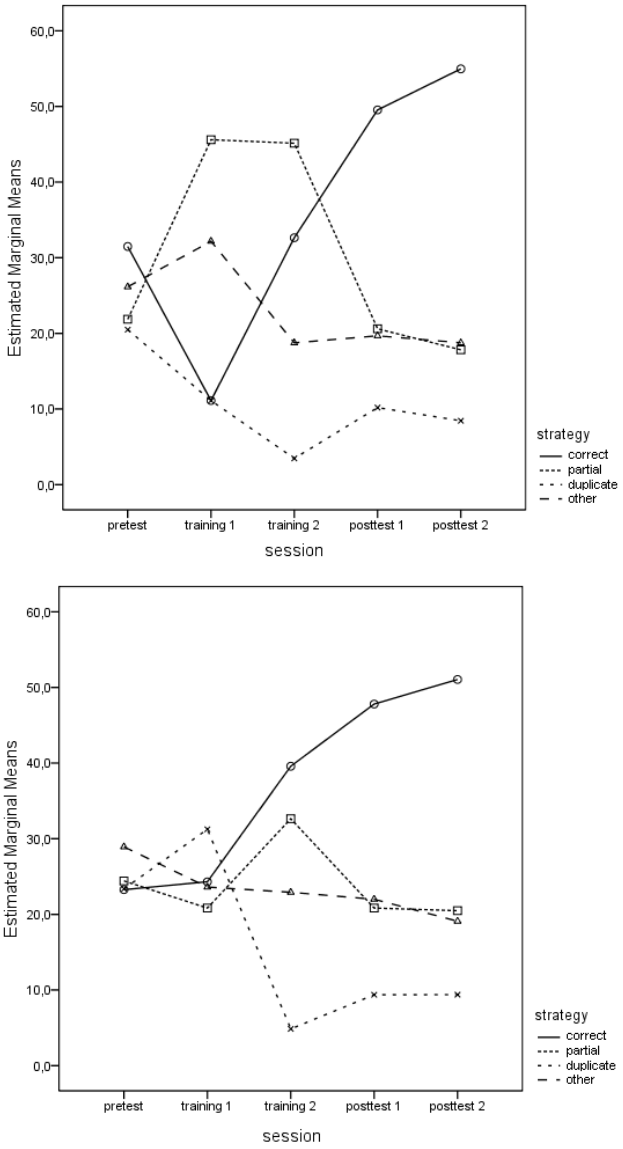

Figure 2.4 *Strategy-use patterns of MC (top) and CR (bottom) trained children.*

2.4 Discussion

The main aim of this study was to investigate the influence of item format on dynamic testing performance of 5-6 year-olds on figural analogical reasoning tasks. The results demonstrate that training in a dynamic testing context with the graduated prompts method leads to greater improvement in analogical reasoning than in untrained controls. No differences in improvement were found between the multiple-choice (MC) and constructed-response (CR) training conditions. However, item format did lead to differences in performance during training. Children trained with CR items provided better quality explanations of analogy solutions compared to those trained with MC items, despite the greater difficulty the children had solving CR items. Also, different strategy-use patterns between the two training groups were found. These results are now discussed in further detail.

As with previous research with dynamic testing and training of analogical reasoning in young children, we found that on the whole the children's ability improved over time, but that training led to greater improvements when compared to untrained controls (e.g., Lifshitz et al., 2005; Tunteler et al., 2008; Siegler & Svetina, 2002). Although we expected that training with CR items would lead to greater progression than training with MC items, the two training conditions did not differ in their improvement after training. On the one hand, one could argue that there is no advantage to CR items, and the advantage in the study of Harpaz-Itay et al. (2006) clearly lies in the construction of the item, indicating that constructed-response may not tap into deeper processing components to the same degree as item construction. On the other hand, any possible advantage in CR may not have been apparent on the MC items of the posttest. For example, Gay (1980) found that when college students were instructed and repeatedly tested in behavioral science knowledge using MC or CR items, no differences were apparent on the MC posttest items, but the advantages of CR training were apparent on

CR posttest items. Including CR items on pre- and posttests in future research could control for this possibility. Furthermore, the items were quite difficult for all participants and the children in the CR-group may have had difficulty transferring their developing skills to a different problem format. Generally, children only show knowledge transfer once they have mastered the correct strategies to solve a task (Siegler, 2006). Nevertheless, despite the posttest advantage for the MC-trained children; they did not perform better than the CR-trained children.

Interestingly, when performance during the training sessions is analyzed, differences between the two training groups emerge. Here we found that CR-trained children provided better quality explanations of how the analogy was solved compared to MC-trained children. Training with CR may lead to a better understanding of analogical reasoning; however further research is needed. Possibly including items or questions in the posttest that require more active processing, such as self-explanation or item construction, would provide the children with more opportunity to demonstrate the depth of their understanding. For example, presenting an analogy construction task in a reversal situation stimulates active processing by asking the child to be the teacher and explain his or her constructed problem thereby providing additional diagnostic information (Bosma & Resing, 2006), and we therefore recommend its use when assessing mastery and understanding of analogical reasoning in future dynamic testing studies.

As with previous research we found that CR items were more difficult than MC items (e.g., Behuniak et al., 1996; In'nami & Kozumi, 2009; Martinez, 1999); the children in the CR condition required more prompts to solve these items and applied fewer correct strategies during training compared to the MC condition. Interestingly, the erroneous strategy used most often by the CR-group was partially correct, rather than duplication as was the case with the MC-group. Duplication is

the most common non-analogical strategy used by young children on classical visual analogies (e.g., Cheshire et al., 2005; Siegler & Svetina, 2002); however analogy strategy-use is most often assessed with MC items. The erroneous strategy-use of the CR training condition, more partial rather than duplication strategies, shows that these children had a good understanding of the required strategy, but made mistakes – forgetting to process one transformation. In other research with CR items partial strategies increase with practice (Stevenson et al., 2011) and training (Tunteler et al., 2008; Tunteler & Resing, 2010; Resing & Elliott, 2011). Perhaps training, especially with CR items, encourages the transition from non-analogical to analogical solutions – albeit incomplete/partial solutions in which one or two transformations are missing. These partial strategies, which could be referred to as utilization deficiencies (Miller & Seier, 1994) of the correct analogy strategy, are most likely due to working memory constraints – a well known bottleneck in young children's analogical reasoning (Richland et al., 2006; Tunteler & Resing, 2010). Another factor that may play a role in the increased partial rather than duplicate solutions for the CR-group is the absence of distractors. These young children may know that duplication is not the solution of how to solve the analogies, but are unable to inhibit responses to distractors leading to relatively more duplication errors, as was the case during training for our MC-group. Inhibition control has been found to play a role in analogy solving in young children (Richland et al., 2006; Thibaut et al., 2008) and future research should investigate whether this is also the case with CR analogies. After training, the children in this study generally showed significant improvement in correct analogical reasoning and training may therefore help children inhibit non-analogical responses (e.g., Siegler & Svetina, 2002; Tunteler & Resing, 2010). On the whole differences in strategy-use between the conditions were not present on the MC items before or after training. Future research into the effects of item format on strategy-use and the possible interaction

with executive functioning, particularly working memory, may provide further insights into the development of analogical reasoning in children.

In sum, CR items may improve learning and provide more fine-grained analysis of strategy-use and are therefore deemed useful in the dynamic testing context. The possible diagnostic advantages of CR items were not examined in this study, but given its relevance for dynamic testing, we recommend future research to investigate this. CR items may be very beneficial for process-oriented diagnostics, with the goal of adapting instruction to individual needs where the analysis of strategy progression and extent of understanding are of particular interest (e.g., Grigorenko, 2009a; Jeltova et al., 2007).

CHAPTER 3

Dynamic testing of ethnic minority children's potential for learning to solve analogies

This chapter is based on Stevenson, C. E., Heiser, W. J. & Resing, W. C. M. (under review). Dynamic testing of ethnic minority children's potential for learning to solve analogies.

Abstract

Dynamic testing is a method to assess cognitive potential in which training is incorporated into the assessment process. This type of assessment appears especially effective for disadvantaged populations such as ethnic minorities or learning disabled children. In this study we present a dynamic test of figural analogy matrices utilizing graduated prompting techniques. We investigate whether the dynamic test outcomes are moderated by ethnicity by comparing the progression of dynamically tested children (n=111) with a practice and an attention control group at three inner-city schools with culturally diverse populations. The results showed that children trained with graduated prompting progressed more quickly in analogical reasoning than both control groups. Cultural background (dominant versus minority culture) was related to initial performance, but not performance gain. The number and type of prompts required during training provided further information on the children's potential for learning and instructional-needs. These were related to pretest performance, performance gain, teacher ratings of learning ability and working memory capacity, but not cultural background. We conclude that graduated prompting of figural analogies has potential as a multicultural dynamic assessment instrument, but this must be demonstrated by assessing its predictive and prescriptive value in culturally diverse groups.

Acknowledgments

We would like to thank Astrid Slingerland for her help with the development of the material and also Arno Martha, Hatice Uysal and Paul Manning for their assistance with data collection. We also would like to thank Paul De Boeck and Marian Hickendorff for their help with the initial IRT analyses.

3.1 Introduction

Dynamic testing can be defined as a method to evaluate cognitive potential that goes beyond traditional assessment approaches by providing information on one's ability to learn from instruction and feedback interventions during the assessment process (Elliott et al., 2010). Dynamic testing is often contrasted with static testing, of which traditional IQ tests are a typical example. Educational and school psychologists often use conventional, static tests in their daily practice, given that cognitive assessment scores are good predictive measures of school achievement and can be input for diagnoses of learning difficulties (Resing, 1997). Yet, critics argue that conventional tests are not the best instruments for determining learning efficiency, as they measure previous learning rather than ability to profit from instruction, can underestimate cognitive ability – especially in disadvantaged groups such as ethnic minorities or learning disabled, and do not provide substantial prescriptive diagnostic information (Elliott, 2003; Fabio, 2005; Grigorenko, 2009a; Haywood & Lidz, 2007). Dynamic tests can provide useful information for educational psychologists with regard to individual differences in learning and potential or instructional-needs (e.g., Bosma & Resing, 2012; Resing et al., 2009). The aim of this paper is to investigate whether similar indices of potential for learning, such as performance change and instructional-needs, can be found in both indigenous and ethnic minority children on a dynamic test of analogical reasoning.

Dynamic testing is considered a promising method for multicultural assessment (Grigorenko, 2009b; Sternberg et al., 2007). This is in contrast with findings of static assessment such as with intelligence or scholastic achievement tests which have been criticized for cultural bias as the dominant culture group generally obtains higher scores (e.g., Fagan & Holland, 2007; Freedle, 2003; Helms-Lorenz & Van de Vijver, 1995). Cultural bias can stem from the tests themselves (i.e. item bias), the testing situation (e.g. nonnative instruction language, cultural influences on

test-wiseness) or cultural differences in the tested construct, such less value being placed on the measured construct (Van de Vijver & Poortinga, 1997; Sternberg et al., 2002; Sternberg, 2004). However, dynamic testing methods may reduce cultural bias as repeated testing or training may compensate for differences in factors such as amount of learning opportunities, test-wiseness or non-native instruction language provide disadvantaged children more opportunity to reveal their cognitive potential (e.g., Bridgeman & Buttram, 1975; Sternberg et al., 2002; Van de Vijver, 2008). For example, Pena, Iglasius, and Lidz (2001) found that dynamic measures of word-learning were better able to distinguish between typically developing and low language ability children than static measures in young children from a culturally diverse population. Similarly, Hamers, Hessels, and Pennings (1996) demonstrate that the evaluation of test scores on a dynamic intelligence test showed that 25-30% fewer ethnic minority children would be categorized as intellectually disabled when using dynamic rather than traditional test scores, whereas only a small percentage of the indigenous Dutch children's categorization would change. Tzuriel and Kaufman (1999) also reported advantages in the dynamic assessment of Ethiopian immigrant children who improved more than their native Israeli counterparts from the mediational process.

In this study we examine whether two indices of potential for learning, performance change and instructional-needs, differ between indigenous Dutch and ethnic minority children. Previous studies have demonstrated that ethnic minority children can "close the gap" in performance with indigenous peers when given sufficient training in the form of dynamic assessment (e.g., Sternberg et al., 2007; Tzuriel & Kaufman, 1999). However, one of the reasons that dynamic assessment procedures are not often used in practice is that these are often time consuming (Grigorenko & Sternberg, 1998). Yet a dynamic approach to the assessment of ethnic minority children seems advisable (Sternberg et al., 2007). In this study we

investigate whether a dynamic test with a short intervention procedure can still provide reduced cultural bias in learning potential indicators for ethnic minority children (e.g., Hessels, 2000). The underlying principle of a standardized dynamic test, in contrast with more extensive dynamic assessment, is not to bring about lasting change, but to measure potential for learning using a short dynamic testing procedure (Resing, Elliott, & Grigorenko, 2012).

Graduated prompting is a specific form of training used in dynamic testing in which increasingly elaborate feedback is provided - initially stimulating metacognitive skills, then explicitly teaching solution strategies (e.g., Campione & Brown, 1987; Resing & Elliott, 2011). By only providing prompts when the student is unable to solve the task independently, insights into learning efficiency and instructional-needs are obtained (Bosma & Resing, 2012; Bosma et al., submitted). For example, the number of prompts required provides an indication of the amount of instruction a child needs to reach a potential performance level (e.g., Ferrara et al., 1986; Resing, 1997). The type of prompts that best lead to solution may guide choices of the most appropriate classroom instructions or interventions for a particular child (e.g., Bosma & Resing, 2012; Resing, Xenidou-Dervou, Steijn, & Elliott, 2012). Resing et al. (2009) demonstrated that ethnic minority children had different instructional-needs, requiring more cognitive prompts – explaining task-specific problem solving steps – compared to indigenous Dutch children when dynamically tested on a seriation task. Graduated prompting has been utilized for inductive reasoning tasks such as geometric matrices (e.g., Ferrara et al., 1986), verbal analogies and visual exclusion (e.g., Resing, 1990) and seriation (e.g., Bosma et al., submitted; Ferrara et al., 1986; Resing & Elliott, 2011; Resing, Xenidou-Dervou, et al., 2012). In the current study, graduated prompting techniques were adapted to a different inductive reasoning task: figural analogy matrices.

Figural analogies are considered a relatively culture-fair inductive reasoning task

(e.g., Cattell, 1979) and central to intelligence (Carpenter et al., 1990), but they are also assumed to be strongly related to working memory capacity (Beunher, Krumm, & Pick, 2005; Süb, Oberauer, Wittmann, Wilhelm, & Schulze, 2002). Working memory capacity (WMC) measures generally show little cultural bias (e.g., Hedden et al., 2002). However, WMC is related to inductive reasoning ability (e.g., Kyllonen & Christal, 1990) and the development of analogical reasoning (Kail, 2007; Richland et al., 2006; Tunteler et al., 2008) and therefore a possible source of individual differences in the dynamic assessment of these skills. For this reason we examined the efficacy of our dynamic test of figural analogical reasoning in a culturally diverse setting while taking individual differences in working memory into account.

A main difference between the present study and previous studies of dynamic test performance in ethnic minority children is the use of Rasch-scaling for the dynamic test scores of performance change, often referred to as gain scores, from pretest to posttest. Rasch models fall under item response theory (IRT) in which the chance that a person solves an item correctly is modelled based not only on the person's ability, but also on the difficulty of the item (e.g., Embretson & Reise, 2000). IRT models provide important advantages for dynamic testing because we look at change in ability over time and when this is measured with classical test theory (CTT) (e.g., comparison of proportion correct) the gain scores pose some statistical pitfalls (Embretson, 1991b; Von Davier, Xu, & Carstensen, 2010). For example, the classical gain score (posttest correct minus pretest correct) is considered unreliable. Furthermore, the meaning of the gain score can depend on pretest performance; for example, a gain of four correct solutions can mean something different if the child had only one item correct on the pretest than if sixteen were solved correctly. Despite these psychometric disadvantages, an individual's gain from pretest to posttest appears to be a meaningful construct in the dynamic testing context (e.g., Calero, Belen, & Robles, 2011; Embretson & Prenovorst, 2000; Grigorenko & Sternberg,

1998). In the present study we wish to examine whether there are group differences in gain between indigenous and ethnic minority children on our dynamic test; therefore, we include gain scores but estimate these using IRT models given the more favorable reliability of IRT gain scores and their interpretation.

The main aim of the current study was to determine whether dynamic testing of analogical reasoning using a short graduated prompts intervention is able to reduce the effect of a non-Dutch cultural background on the learning potential indices of performance change and instructional-needs. We expected (1) initial differences in (pretest) performance between indigenous Dutch and ethnic minority pupils in figural analogy solving (e.g., Helms-Lorenz & Van de Vijver, 1995; Fagan & Holland, 2007). With regard to potential for learning we hypothesized (2a) commensurate ability of both ethnic minority and indigenous Dutch children to improve (i.e. gain) from pretest to posttest (Tzuriel & Kaufman, 1999; Wiedl, Kampling, Köning, Schrevels, & Waldorf, 2011), but (2b) that these groups would have different instructional-needs during dynamic testing, where ethnic minority children would require more training compared to the indigenous Dutch children (e.g., Hamers et al., 1996; Hessels, 2000). More specifically, greater cognitive prompting needs were expected for ethnic minority children whereas indigenous Dutch children were expected to require more metacognitive prompts (Resing et al., 2009).

3.2 Method

3.2.1 Participants

Participants were 111 children (63 boys, 48 girls) from second grade primary schools (M = 8;1, SD = 5 months). Fifty-six children were categorized as indigenous Dutch (both parents with Dutch nationality) and 55 as ethnic minorities (one or both

parents have a non-Dutch nationality). The participants were recruited from three neighboring primary schools of comparable SES located in an inner-city district in the Netherlands. Written informed parental consent was obtained.

3.2.2 Design & Procedure

A pretest-training-posttest control-group design with randomized blocking was employed. Children were randomly blocked into one of three conditions: (1) graduated prompts, (2) practice control and (3) attention control. The blocking was based on scores on the RAKIT subtest visual exclusion (Bleichrodt et al., 1987), ethnicity (indigenous Dutch or ethnic minority), classroom and gender. Prior to the experimental sessions the visual exclusion task, used to measure inductive reasoning, and the WISC-IV subtest Digit Span Backwards (Wechsler, 2003), used to measure WMC (e.g., Süb et al., 2002), were administered. Also, teachers were requested to rate each child's learning ability.

All children were given the pretest and posttest. During the intervention phase trained children received the graduated prompts training, whereas practice control children received the same items without training or feedback and the attention control group was provided with a maze coloring task. All testing sessions took place weekly in a quiet room within the school and the children were individually tested for a total of 75 to 100 minutes. Following each session children were given a sticker for motivation. Qualified graduate students, trained in advance in all standardized testing and training procedures, administered the tests.

3.2.3 AnimaLogica

This dynamic test of analogical reasoning was comprised of an introduction task, pretest, training and posttest. The visual analogies are classically presented in 2x2 matrix format (Stevenson et al., 2009). Colored (red, yellow or blue) animal

drawings occupied three squares and the lower right or left quadrant was empty. Children had to infer the relation between two pictures (horizontally or vertically) and apply this to a third picture to solve the analogy (A:B::C:D). Rule-based item generation, where item difficulty can be predicted based on the number of figures and transformation rules applied (e.g., Mulholland, Pellegrino, & Glaser, 1980) was used to develop items of varying difficulty. The six transformation rules used were: (1) animal, (2) color, (3) size, (4) position, (5) orientation and (6) quantity. Animal figures rather than abstract or geometric figures were used in conjunction with familiar transformations in order to meet the requirement of familiar objects and relations deemed essential for successful analogical reasoning in young children (Goswami, 1992). The elements (animals), transformations and colors were selected randomly but constrained to comprise near equal representation of each in the task booklets.

Introduction task

Six items consisting of a pair of animals which differed by one of the six transformations were presented. The children were asked to name the animals and explain what changed; mistakes were corrected using a standardized protocol.

Pretest and posttest

The 22 analogy problems used during the tests were solved by choosing a picture from five alternatives at the bottom of the task. The answer options were systematically constructed: (1) correct answer, (2 & 3) partial answer: missing one transformation, (4 & 5) non-analogical answer: duplication or missing two transformations (see Figure 3.1).

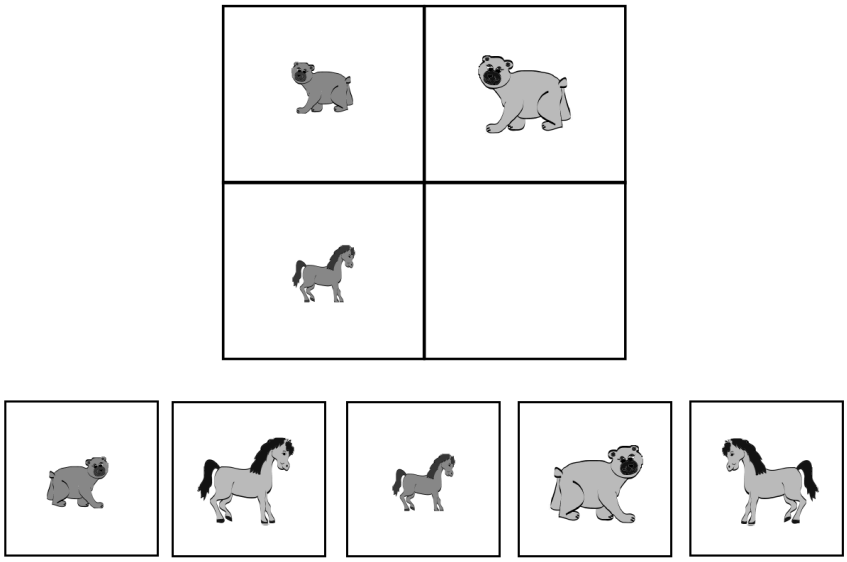

Figure 3.1 *Example MC item from* AnimaLogica *with options representing the strategies (from left to right) non-analogical, correct, duplicate, partial and partial respectively.*

Training items

The 12 training items (see Figure 3.2) were presented in constructed-response format 2. Answers were constructed from a number of animal cards; for each type of animal a box containing plasticized cards of the animal in three different colors and two possible sizes was available. By turning the animal card over the animal's orientation could be changed and the position was altered by moving the card to a different location in the empty quadrant.

Graduated Prompting Procedure

The graduated prompts training phase consisted of small structured steps, ranging from very general to task specific instructions. Each session began with two

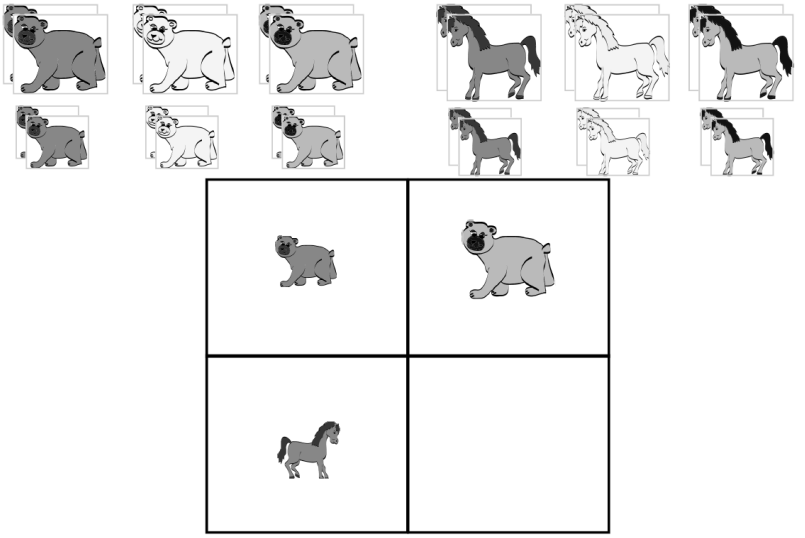

FIGURE 3.2 *Example CR item from* ANIMALOGICA.

example problems, after which the presentation of each item began with a general instruction. The child responded by constructing his/her response with the animal cards. The tester provided feedback on the response. If the answer was correct, the examiner asked the child to explain his/her reasoning before continuing with the next item. If the child's response was not correct a prompt was provided. A cycle of response-prompt-response was repeated until the child constructed the correct answer or the fifth and final prompt had been given (see Figure 3.3). Before continuing with the next item the child was always asked to explain the solution – stimulating learning through self-explanation (e.g., Siegler, 2002; Stevenson et al., 2009). Both textual and pictorial descriptions were included in the prompts and the instructions were emphasized with gestures to provide extra support for differences in language abilities.

As with the Learning potential of Inductive Reasoning test (Resing, 1990), the test upon which AnimaLogica is based, the first two prompts focus on metacognitive skills emphasized in cognitive training studies (e.g., Campione & Brown, 1987; Schraw, 1998). The third to fifth prompt focus on the cognitive process of solving the analogy based on Sternberg's (1977) basic cognitive processes of analogical reasoning: encoding, inference, mapping, application, comparison, justification and response. The first prompt aided problem recognition and redefinition, where the child was asked how such an item was solved before and was provided with more detailed instruction. In the second prompt the aid card was given, which presented the general steps to solve the analogies (see Figure 3.4). In the third prompt, guiding encoding and inference, the examiner worked through the steps on the aid card explaining with both words and gestures. For example, what changes from here to here (A:B)? In the fourth prompt the horizontal and vertical transformations were summarized and the inference and mapping steps were emphasized. In the final prompt the examiner used scaffolds to help the child systematically solve the problem per transformation, such as "Which animals belong in the empty box?", "What color should the elephant be?", "Which direction should the dog face?". After each question direct feedback was given, guiding the child step-by-step to the correct solution.

3.2.4 Scoring

The children's answers to the analogy problems were based on the selected or construction answer and scored as correct/incorrect. For the pretest and posttest Rasch estimates from item response theory (IRT, e.g., Embretson & Reise, 2000) were calculated to determine initial ability (pretest performance) and potential ability (posttest performance). In IRT examinee ability is modeled using both the responses per item and item properties such as difficulty level. We applied Andersen's Rasch

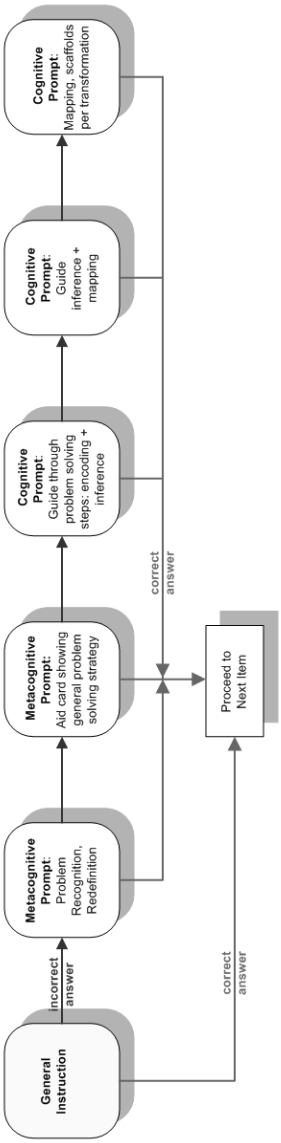

FIGURE 3.3 *Flowchart of graduated prompting procedure.*

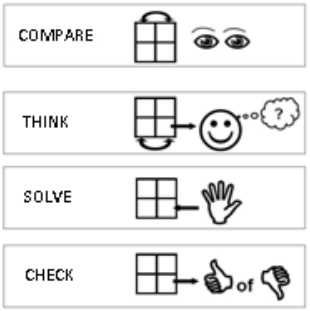

Figure 3.4 *Aid card used during graduated prompting.*

model for repeated measurements (Andersen, 1985), which corrects for within child correlations, to estimate the children's pretest and posttest scores and analyze the children's differences in gain between these two measurement moments.

During training the child's solution was recorded after each prompt and prompting of an item ceased if the solution was correct. The sum total of the number of prompts required per item was used to determine the amount of help required to complete the training – with more prompts indicating more difficulty with item solutions. Second, the prompts were categorized as metacognitive (prompts 1-2) or cognitive (prompts 3-5) (see Resing et al., 2009). These were used to investigate patterns of differences in individual needs for instruction.

3.3 Results

Before conducting analyses to answer the research questions we first checked whether the children in the three conditions differed in cognitive functioning or age prior to testing. Furthermore, we describe the psychometric properties of the Rasch-scaled test and items, including analyses of whether item bias occurs for indigenous or ethnic minority children.

3.3.1 Initial Group Comparisons

The children's initial level of inductive reasoning, measured with the visual exclusion task, did not differ between the three conditions ($F(2, 108) = .06$, $p = .94$). There were also no differences in working memory capacity (WMC), teacher rating of learning ability or age between the conditions (see Table 3.1). Differences between indigenous and ethnic minority children on these variables were also examined (for basic statistics see Table 3.2). There were significant differences in mean standardized scores on the visual exclusion task ($F(1, 109) = 6.79$), $p = .01$), with native Dutch children performing better than ethnic minorities. Also teacher ratings of learning ability differed slightly between the groups, whereby ethnic minority children received lower ratings than native Dutch children ($F(1, 93) = 4.63$, $p = .03$, $\eta_p^2 = .05$). There were no significant differences between the ethnic groups with regard to WMC or age in this sample (see Table 3.2).

3.3.2 Psychometric Properties

Cronbach's alpha coefficient of internal consistency for the pretest, $\alpha = .60$, is moderate. For the posttest the reliability was $\alpha = .80$ and is considered good. The pretest proportion correct responses per item ranged from .14 to .95 and for the posttest from .21 to .98. The rank correlation between the proportion incorrect and the predicted difficulty level based on the number of transformations was $\rho = .75$, $p < .001$ for the pretest and $\rho = .76$, $p < .001$ for the posttest. The correlation of the pretest and posttest proportion correct across individuals was $r = .50$, $p < .001$.

The independent Rasch (1 PL) model parameters were estimated for the pretest and posttest using the Marginal Maximum Likelihood (MML) estimator in the `ltm` package for R (Rizopoulos, 2006). A parametric Bootstrap goodness-of-fit test using Pearson's χ^2 statistic was used to investigate model fit. Based on 50 datasets the Rasch model fit of the pretest was slightly deviant and the posttest was acceptable

Table 3.1 *Means and standard deviations on visual exclusion, working memory, teacher rating of learning ability and age per condition (graduated prompts, practice control and control).*

Variables	N	M	SD
Visual Exclusion[1]			
Graduated prompts	37	16.16	5.19
Practice Control	37	16.19	5.16
Control	37	15.81	5.16
Total	111	16.05	5.13
Working Memory[2]			
Graduated prompts	37	6.14	1.25
Practice Control	37	6.03	1.26
Control	37	6.27	1.17
Total	111	6.14	1.22
Teacher Ratings of Learning Ability[3]			
Graduated prompts	37	4.24	0.80
Practice Control	37	4.41	0.69
Control	37	4.27	0.96
Total	111	4.31	0.82
Age[4]			
Graduated prompts	37	98.41	5.45
Practice Control	37	96.68	5.02
Control	37	96.62	4.39
Total	111	97.23	5.00

[1] based on standardized scores of the RAKIT visual exclusion subtest
[2] sum score on WISC-IV digit span backwards
[3] teachers scored children's learning ability on a sliding scale of 0 (very low learning ability) - 7 (very high learning ability)
[4] in months

TABLE 3.2 *Means and standard deviations on visual exclusion, working memory, teacher rating of learning ability and age per ethnic group (indigenous Dutch or ethnic-minority).*

	N	M	SD
Visual Exclusion[1]			
indigenous	56	17.30	5.20
ethnic minority	55	14.78	4.76
Digit Span Backwards[2]			
indigenous	56	6.32	1.31
ethnic minority	55	5.96	1.11
Teacher Ratings of Learning Ability[3]			
indigenous	56	4.45	0.76
ethnic minority	55	4.16	0.86
Age[4]			
indigenous	56	96.89	4.72
ethnic minority	55	97.58	5.28

[1] based on standardized scores of the RAKIT visual exclusion subtest
[2] sum score on WISC-IV digit span backwards
[3] teachers scored children's learning ability on a sliding scale of 0 (very low learning ability) - 7 (very high learning ability)
[4] in months

(p = .04 and p = .28 respectively). The correlation between the item parameters of the pretest and posttest was very strong: r = .99. Therefore we considered the application of Andersen's Rasch Model for repeated measurements (Andersen, 1985) appropriate. This was implemented with the `lmer4` package for R (Bates & Maechler, 2010) as described by De Boeck et al. (2011).

Differences in item functioning for the two ethnic groups were investigated for the pretest and posttest responses on the test items. In analyses of differential item functioning (DIF) the probability of a correct response given the same ability level is compared between the two groups (e.g., Facon, Magis, Nuchadee, & Boeck,

2011). The Mantel-Haenszel procedure and Raju's DFIT method (Magis, Béland, Tuerlinckx, & Boeck, 2010) were used to test for uniform DIF. Neither procedure revealed significant differences between ethnicities in functioning for any of the items.

3.3.3 Pretest to posttest progression

Our first research question concerned the effect of the graduated prompts training on the children's progression in analogical reasoning from pretest to posttest and whether this was related to ethnicity. This was investigated using repeated measures (RM) ANOVA with Rasch-scaled ability estimates per session as dependent variable (see bottom of Table 3.3 for basic statistics), with Session as within-subjects variable and Condition and Ethnicity as between-subjects variables and working memory as a covariate. The main effect for Session was significant (Wilks' $\lambda = .63$, $F(1, 105) = 62.10$, $p < .001$, $\eta_p^2 = .37$) showing that children, on average, progressed in figural analogy solving across sessions. The significant interaction effect for Session x Condition (Wilks' $\lambda = .86$, $F(2, 105) = 8.57$, $p < .001$, $\eta_p^2 = .14$) indicates that children in the conditions differed in their degree of progression. As can be seen in Figure 3.5 children in the graduated prompts condition improved more than the practice and the attention control conditions. Interestingly, no significant differences were found between practice and attention control conditions. Large standard deviations were found, most notably on the posttest, indicating great variation in the children's initial ability and their ability to profit from the training or control tasks.

A main effect of Ethnicity was found: $F(1, 104) = 10.93$, $p = .001$, $\eta_p^2 = .10$. As can be seen in the means reported in Table 3.3 indigenous children generally performed better on pretest and posttest measures. More importantly there was no effect of Session x Ethnicity (Wilks' $\lambda = .98$, $F(1, 104) = 2.18$, $p = .14$, $\eta_p^2 = .02$). The interaction Session x Condition x Ethnicity was not significant (Wilks' $\lambda = .99$,

TABLE 3.3 *Means and standard deviations of Rasch-scaled pretest and gain estimates per condition (graduated prompts, practice control and control) and ethnic group (indigenous or ethnic-minority).*

			Pretest		Posttest	
Condition	Ethnicity	N	M	SD	M	SD
Graduated prompts	indigenous	19	.10	.42	.83	.68
	minority	18	-.11	.46	.72	.86
	Total	37	.00	.45	.86	.78
Practice control	indigenous	19	.22	.31	.66	.67
	minority	18	-.19	.40	-.05	.84
	Total	37	.02	.41	.32	.83
Attention control	indigenous	18	.27	.68	.76	1.38
	minority	19	-.14	.34	.04	1.00
	Total	37	.06	.57	.40	1.24
Total	indigenous	56	.19	.49	.81	.95
	minority	55	-.14	.40	.23	.96
	Total	111	.03	.48	.52	1.00

$F(2, 104) = .52$, $p = .60$, $\eta_p^2 = .01$) indicating that the gain from pretest to posttest does not differ between ethnic groups per condition (see Figure 3.6). Indigenous children had higher estimates of initial ability than ethnic minorities which is in line with hypothesis 1. However, no differences in gain after graduated prompting were present between the two groups, confirming hypothesis 2a.

The effect of Working Memory was also significant: $F(1, 104) = 7.96$, $p = .006$, $\eta_p^2 = .07$. Children with higher working memory scores generally had higher pretest and posttest scores ($r = .24$, $p = .01$ and $r = .29$, $p = .01$ respectively). The difference between these correlations is not significant ($z = -.04$, $p = .69$).

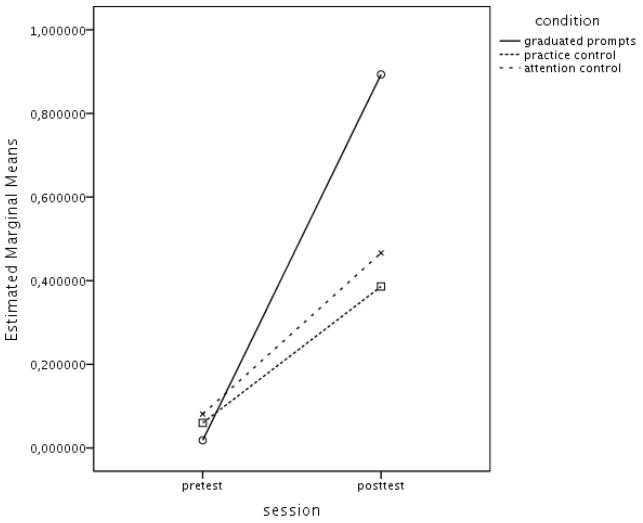

Figure 3.5 *Estimated marginal means of ability per condition across sessions.*

3.3.4 Instructional-needs during training

The number of prompts the children required while solving the training tasks showed great variation ($M = 14.97$, $SD = 8.00$). Although the children's need for prompts significantly lessened from the first ($M = 10.86$, $SD = 5.36$) to second training session ($M = 4.11$, $SD = 3.39$), $F(1, 36) = 103.28$, $p < .001$, $\eta_p^2 = .74$, large variation remained. A univariate ANOVA was conducted to determine whether total number of required prompts (dependent variable) was related to ethnicity (between-subjects factor). This was not the case: $F(2, 105) = .62$, $p = .54$, $\eta_p^2 = .01$. However, as can be seen in Table 3.4 strong Pearson correlations were found between WMC, teacher ratings of learning ability, pretest ability or gain and total number of required prompts.

Prompts were categorized as metacognitive (prompts 12) and cognitive (prompts

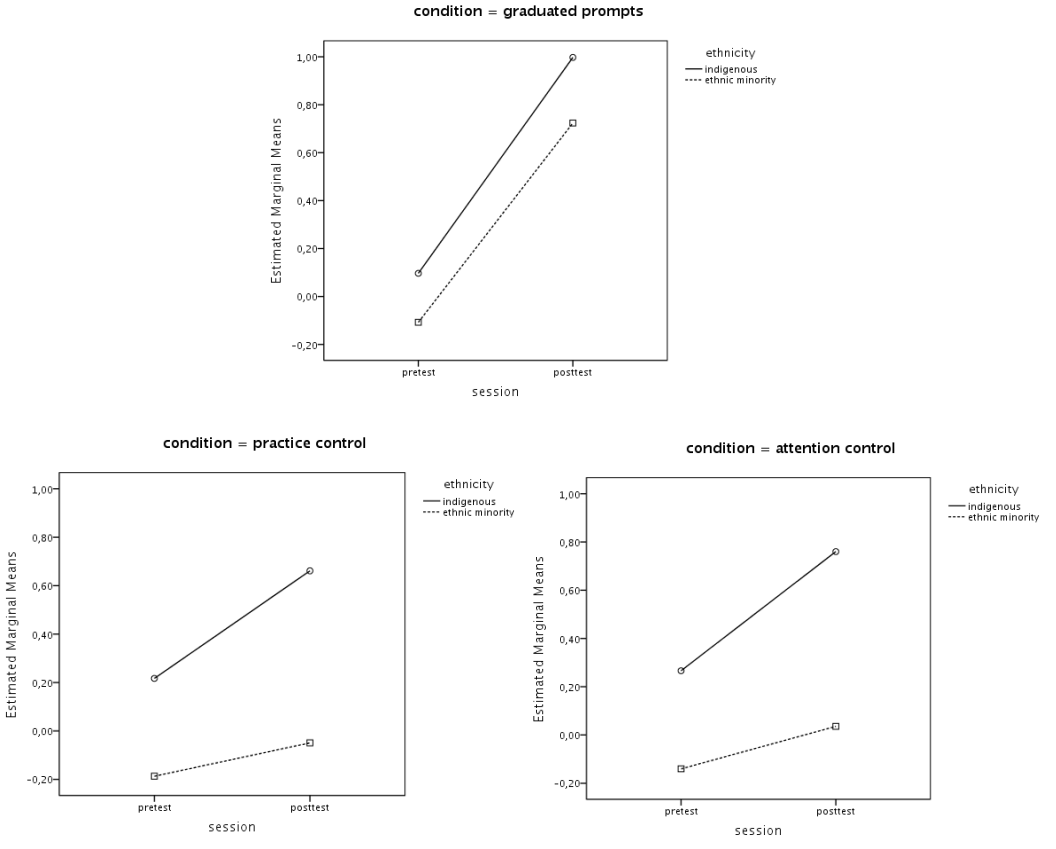

FIGURE 3.6 *Estimated marginal means of ability on pretest and posttest per condition per ethnicity.*

Table 3.4 *Pearson correlations (correlation above diagonal, p-value below diagonal) between total required prompts and learning ability rating by teachers, Digit span backwards, Visual exclusion, pretest ability and gain (posttest – pretest score).*

	Total prompts	Teacher rating	WMC	Visual Exclusion	Pretest	Gain
Total prompts		-.48	-.42	-.55	-.74	-.40
Teacher rating	< .01		.25	.36	.51	-.06
WMC	< .01	.07		.45	.38	.28
Exclusion	< .001	< .001	< .01		.47	.23
Pretest	< .001	< .05	< .01	<.01		.01
Gain	< .01	.36	< .05	.09	.48	

Table 3.5 *Means and standard deviations of number of required prompts per ethnic group (indigenous Dutch or ethnic-minority).*

		Metacognitive prompts		Cognitive prompts	
Ethnicity	N	M	SD	M	SD
Indigenous	19	4.32	1.25	2.16	1.54
Minority	18	3.78	1.59	2.67	1.97
Total	37	4.05	1.43	2.41	1.76

3-5); see Table 3.5 for basic statistics. On the whole fewer cognitive than metacognitive prompts were provided ($F(1,36) = 20.76$, $p < .001$, $\eta_p^2 = .37$). To analyze whether ethnic minority children required more cognitive prompts than indigenous children, a MANOVA with ethnicity as between-subjects factor was conducted. This did not reveal a significant effect for the dependent variables total metacognitive prompts ($F(1,35) = 1.32$, $p = .26$, $\eta_p^2 = .04$) or total cognitive prompts ($F(1,35) = .77$, $p = .39$, $\eta_p^2 = .02$), therefore hypothesis 2b could be rejected. WMC however was linked to instructional-needs as children with greater WMC required fewer cognitive prompts than those with a smaller WMC: the correlations between

WMC and total metacognitive and cognitive prompts were $r = .04$, $p = .804$ and $r = -.46$, $p = .005$ respectively.

3.4 Discussion

Our main findings show that the short graduated prompting procedure is an effective intervention form in the dynamic testing of figural analogical reasoning in both indigenous Dutch and ethnic minority children. As with previous research repeated testing led to spontaneous improvements in analogical reasoning, with more marked improvement due to training, yet great variability in initial ability as well as performance change (e.g., Freund & Holling, 2011a; Siegler & Svetina, 2002; Tunteler et al., 2008). The graduated prompting procedure led to greater improvement from pretest and posttest than repeated testing of both practice and attention-control conditions and was demonstrated to be an effective means of improving analogy solving with significant large effects comparable to that of other dynamic tests, despite the shorter duration (e.g., Resing et al., 2009; Resing & Elliott, 2011).

Ethnicity was found to be related to initial performance on AnimaLogica as indigenous Dutch children obtained on average higher ability estimates on the pretest than ethnic minorities (e.g., Te Nijenhuis & Van Der Vlier, 2001; Van de Vijver, 2002, 2008). This did not appear to be due to cultural bias on the item level as none of the items appeared to function differently for the two ethnic groups. No differences in the Rasch-scaled gain from pretest to posttest were found between indigenous and ethnic minority children, nor were differences present in instructional-needs, i.e. the number and type of required prompts during training. This finding of similar indices for potential for learning between the two ethnic groups coincides with earlier investigations into cultural differences on dynamic tests (Hamers et al., 1996; Sternberg et al., 2007; Tzuriel & Kaufman, 1999) and our findings are

further supported by the culture-fair results of studies that allow for equal learning opportunities of the assessed task prior to testing (e.g., Bridgeman & Buttram, 1975; Fagan & Holland, 2009). However, in contrast to Resing's study (2009) we did not find differential instructional-needs between the indigenous Dutch and ethnic minority children. This may be due to differences in the assessed task; perhaps figural matrices have a lower cultural loading than mathematical seriation problems (Helms-Lorenz, Van de Vijver, & Poortinga, 2003) leading to fewer differences in instruction. Yet, cultural bias in figural analogical reasoning may still be present when ability is interpreted in the traditional sense as ethnic minorities appear to have systematically lower pretest scores (Sternberg et al., 2007; Van de Vijver, 2008). However, dynamic measures, quantified by gain and required instruction, do not appear to suffer from this bias – neither in longer assessments (e.g., Sternberg et al., 2007; Tzuriel & Kaufman, 1999) nor in short-term measures applying graduated prompting techniques, such as AnimaLogica. Furthermore, these findings appear consistent for various forms of inductive reasoning tasks such as figural analogies and seriation (e.g., Hessels, 2000; Resing et al., 2009; Tzuriel & Kaufman, 1999). Our results add to the building body of evidence that dynamic testing has potential as a form culture-fair assessment of multicultural groups. Future investigations should examine the topics of cultural bias and equivalence of learning potential indicators of dynamic tests utilizing graduated prompting procedures in other areas such as reading and math with a larger sample in more depth. Furthermore, particular attention should be paid to the utility of the gain score – and the value of using IRT-scaling to reliably estimate children's performance change during dynamic testing.

Working memory was included in our analyses given its strong relation to figural matrix solving capacity (Kyllonen & Christal, 1990; Kail, 2007). We found that working memory was related to the children's initial ability on the figural analogies

in AnimaLogica, whereby children with greater working memory efficiency had higher ability estimates. This indicates that in future investigations of group differences, such as ethnicity, in AnimaLogica, and perhaps other dynamic tests of analogical reasoning, it is important to control for the effects of individual differences in working memory capacity. Performance change was not moderated by working memory. Yet, the amount of instruction required during graduated prompting to solve the analogy tasks correctly was related to a number of the measured factors, of which WMC is just one. For example, teacher ratings of learning ability correlated strongly with required prompts. It appears we are tapping into similar information the teacher obtains in the classroom on individual children's ability to learn from instruction (Bosma & Resing, 2012). Furthermore, the student's gain from pretest to posttest was related to the amount of required instruction, where children requiring fewer prompts generally improved more. However, as with previous dynamic tests, there is much variability in the children's performance and instructional-needs (e.g., Bosma et al., submitted; Resing et al., 2009). The importance of the required prompts lies in providing profiles of an individual's instructional-needs (Bosma & Resing, 2012; Bosma et al., submitted) and translating this to information that is useful for classroom instruction (e.g., Jeltova et al., 2007). Investigating the usefulness of instructional-needs based on the number and type of prompts should be a focus of future research with AnimaLogica in order to effectively determine its worth for psychoeducational practice.

In sum, although group differences between ethnic minority and indigenous children occur in initial ability to solve analogies, their outcomes on the dynamic measures of performance change and instructional-needs were similar. Working memory capacity does not appear to influence ability to profit from the graduated prompts training; it does play a role in both solution ability and instructional-needs and therefore requires careful investigation in future studies. The relevance of

the dynamic outcomes for psychoeducational practice were only briefly touched upon in this initial study, however we recommend future research to focus on the predictive and prescriptive diagnostic value of this test in culturally diverse groups to further investigate the potential of AnimaLogica for multicultural assessment.

CHAPTER 4

Prompting learning and transfer of analogical reasoning: Is working memory a piece of the puzzle?

This chapter is based on Stevenson, C. E., Heiser, W. J. & Resing, W. C. M. (under review). Prompting learning and transfer of analogical reasoning: Is working memory a piece of the puzzle?.

Abstract

Dynamic testing is an assessment approach that aims to assess potential for learning by measuring performance improvement as a response to training while testing. In this study we use this approach in order to: (1) determine whether training children in analogical reasoning affects transfer of inductive reasoning skills to other tasks and (2) explore the relationship between working memory, training and transfer effects. This was investigated using a pretest-training-posttest control group design with 64 participants, aged 7-8 years (M = 7.6 years; SD = 4.7 months). All of the children were tested on four inductive reasoning tasks. Half of the children were trained in solving figural analogies according to the graduated prompts method, while the control group practiced with these items. Initial ability and performance change from pretest to posttest were estimated using Embretson's (1991b) item response theory Multidimensional Rasch Model of Learning and Change. We found that the short training procedure improved figural analogical reasoning more than practice. Working memory was strongly related to initial performance on each of the inductive reasoning tasks. Yet, we found that performance change and knowledge transfer were only somewhat related to initial ability and unrelated to working memory. This indicates that performance change and ability to transfer trained skills to new tasks may be separate constructs and of possible importance in the assessment of learning and cognitive potential.

Acknowledgments

We would like to thank Hester van den Akker and Colette Kuijpers for their assistance with data collection and coding and Hester for her additional contribution to data analysis.

4.1 Introduction

Dynamic testing can be defined as a diagnostic method that focuses on potential for learning and aims to provide insight into developing abilities (Elliott, 2003; Sternberg & Grigorenko, 2002). Dynamic assessment diverges from traditional assessment in that feedback is provided by the examiner during testing in order to facilitate learning and gain insight into learning efficiency and cognitive potential (Elliott et al., 2010). In dynamic testing, various indices are used to examine a child's potential for learning, such as performance improvement following feedback interventions (e.g., Hessels, 2009; Tzuriel, 2001), the amount and type of instruction that best aides task solution (e.g., Bosma & Resing, 2012; Resing & Elliott, 2011), and ability to transfer these newly developed skills to other problems (Campione & Brown, 1987; Day et al., 1997; Lidz & Pena, 1996; Resing, 1997; Sternberg & Grigorenko, 2002). Previous research demonstrates that in dynamic testing designs using a pretest-training-posttest format the interventions generally lead to improve an examinee's ability in the assessed skill (e.g., Day et al., 1997; Sternberg & Grigorenko, 2002). Furthermore, graduated prompting, a specific form of intervention, can provide insight into the examinee's instructional needs (e.g., Bosma et al., submitted; Resing, Xenidou-Dervou, et al., 2012). In earlier dynamic testing research utilizing graduated prompting techniques the ability to transfer what was learned during the intervention was sometimes included in the assessment process (Brown & Kane, 1988; Campione, Brown, Ferrara, Jones, & Steinberg, 1985; Ferrara et al., 1986; Resing, 1993). However, transfer measures have received less attention in the more recent literature, perhaps due to the difficulty in eliciting transfer (e.g., Barnett & Ceci, 2002; Bransford & Schwartz, 1999; Detterman, 1993; Hager & Hasselhorn, 1998; Roth-Van Der Werf, Resing, & Slenders, 2002). Yet, transfer of skills to novel situations may provide insights into a child's potential for learning (e.g., Bosma & Resing, 2006; Ferrara et al., 1986). In the present study we investigated the extent to

which reasoning skills learned during the dynamic testing of analogical reasoning were applied to similar untrained tasks. Furthermore, because inductive reasoning and working memory capacity (WMC) appear to be inter-related in children (e.g., Kail, 2007) we investigated the role of WMC on the near-transfer of inductive reasoning skills in a dynamic testing situation.

4.1.1 Dynamic testing of inductive reasoning

Dynamic tests often include inductive reasoning tasks (e.g., Ferrara et al., 1986; Resing & Elliott, 2011; Resing et al., 2009), which are considered central to intelligence (Carpenter et al., 1990; Klauer & Phye, 2008). Analogical reasoning, a form of inductive reasoning, is deemed essential to school learning and refers to the capacity to learn about a new situation by relating it to a structurally similar more familiar one (Goswami, 1992). Classical analogies (A:B::C:?) and figural matrices (see Figure 4.1) are often included in measures of cognitive ability and considered strongly related to 'g' (Freund & Holling, 2011a; Primi, 2001). The ability to reason by analogy is assumed to develop with great variability throughout childhood (e.g., Leech et al., 2008; Siegler & Svetina, 2002; Tunteler & Resing, 2007a). Older children tend to perform better than younger children, which may be explained by improvements in efficiency of working memory capacity (Kail, 2007; Richland et al., 2006). Improvement in analogical reasoning can take place spontaneously with practice (e.g., Tunteler & Resing, 2002), with further learning effects provided by feedback (Cheshire et al., 2005), self-explanation (Siegler & Svetina, 2002; Stevenson et al., 2009) and other training formats (e.g., Alexander, Willson, White, & Fuqua, 1987; Klauer & Phye, 2008; Tunteler et al., 2008). Training with graduated prompting techniques has been shown more effective than practice alone with regard to both learning and transfer (Bosma & Resing, 2006; Ferrara et al., 1986; Tunteler & Resing, 2010). Training type may also play a role in the learning and transfer of analogical

reasoning (e.g., Harpaz-Itay et al., 2006; Stevenson, Heiser, & Resing, under review).

4.1.2 Transfer of inductive reasoning skills

The ability to spontaneously generalize a problem-solving approach taught in one context to a different but applicable situation is referred to as transfer. This is considered an important aim of formal schooling (e.g., De Corte, 2003). However, numerous studies show that transfer doesn't occur easily as learning is context-bound and children rarely recognize that their acquired problem solving skills can be applied in novel situations (e.g., Barnett & Ceci, 2002; Bransford & Schwartz, 1999; Luo, Thompson, & Detterman, 2003; Siegler, 2006). Transfer can be assessed broadly such as from school learning to real-life situations or in a more narrow manner – from one cognitive task to a structurally similar one, referred to as near-transfer. Near-transfer also appears not to be common-place (see Jacobs and Vandeventer (1971) for this distinction). For example, Roth-Van der Werff et al. (2002) systematically assessed whether children trained in solving inductive reasoning tasks were able to generalize the learned problem solving skills to superficially similar and dissimilar problems that measured the same inductive reasoning skills. In their study, the children improved more on superficially similar tasks than those who only practiced with the same items. Yet, changes on superficially dissimilar tasks could be attributed to practice effects.

However, children may show greater transfer of knowledge when the targeted strategy has been mastered (Siegler, 2006). For example, Tunteler & Resing (2010) found that 8-year-olds who obtained high scores on a geometric analogy task improved more on a verbal analogies near-transfer task during the posttest. But as with Roth-Van der Werff et al. (2002) the improvement on the superficially dissimilar verbal analogy task in Tuntler & Resing's study was independent of having received training – practice alone appeared to elicit transfer in high ability

children. Aside from practice effects, instructional conditions also appear to play a role in near-transfer. For example, Harpaz-Itay, et al. (2006) found that 12-year-olds trained in verbal analogy solving also improved on geometric and numerical analogies, however the transfer effects were greater in children trained in analogy construction as opposed to multiple-choice solution.

In this study children were either trained to solve figural analogies in constructed-response format or practiced with these items (e.g., Stevenson et al., under review; Stevenson, Heiser, & Resing, submitted 2011a). We investigated whether training with graduated prompting or initial ability level played a role in the transfer of inductive solving skills to three related inductive reasoning tasks differing in content, format and/or measured construct. First, the geometric analogies used by Tunteler & Resing (2010), which differed only in content from the dynamically tested figural analogies. Second, an analogy construction task (e.g., Harpaz-Itay et al., 2006) in a form for younger children where roles of examiner and child are reversed (Bosma & Resing, 2006), which differed in format but not content or measured construct. Finally, a geometric and numerical seriation task (Durost, Gardner, & Madden, 1970), also included in Roth-Van der Werff et al.'s study (2002), was used that differed in content and construct (i.e. series completion rather than analogical reasoning).

4.1.3 Working memory and inductive reasoning

The influence of working memory capacity on the training and transfer of inductive reasoning in a dynamic testing context requires further research given that many researchers have found a strong relationship between working memory capacity (WMC) and inductive reasoning ability (e.g., Bacon, Handley, Dennis, & Newstead, 2008; Conway, Cowan, Bunting, Therriault, & Minkoff, 2002; Krumm & Buehner, 2008; Kyllonen & Christal, 1990; Morrison, Holyoak, & Truong, 2001; Süb et al.,

2002). Baddeley and Hitch (1974) proposed a model to describe the structure of WMC in which the central executive system is considered responsible for controlling attention and information processing, which regulates the operation of two domain-specific systems, the phonological loop and visuospatial sketchpad. The structure described by the Baddeley & Hitch model appears present and assessable in young children (Alloway, Gathercole, & Pickering, 2006; Gathercole, Pickering, & Wearing, 2004; Swanson, 2008) and related to young children's analogical reasoning (e.g., Cho, Holyoak, & Cannon, 2007; Kail, 2007; Richland et al., 2006). For example, Krumm et al. (2008) found that working memory predicts a large amount of variance in reasoning ability. Furthermore, significant relations have been found between increases in efficiency of working memory capacity (WMC) and increases in reasoning and problem solving (Kail, 2007; Swanson, 2008). Tunteler & Resing (2010) found that memory of abstract figures was related to performance on the geometric analogies task, included as a transfer task in this study. Richland et al. (2006) found that children's ability to solve scene analogies was related to their performance on a verbal WMC task. The separate contribution of the verbal and visuospatial components to figural analogy matrices utilized in the present study has not yet been investigated. We therefore extend the work of previous studies by including measures of both verbal and visuospatial WMC.

Working memory may become more efficient due to training and this automation of skills may result in greater transfer effects (Dahlin, Neely, Larsson, Bäckmann, & Nyberg, 2008; Jaeggi, Buschkuehl, J., & Perrig, 2008). It is plausible that training during dynamic assessment may lead to performance change and transfer effects through similar mechanisms. For example, we found that children's WMC was related to improvement in analogy solving in untrained children but not in children who received training with graduated prompting techniques (Stevenson et al., submitted 2011a). Similarly, in a dynamic test utilizing the inductive reasoning task

seriation, children with lower verbal WMC scores improved comparably to those with greater WMC scores but the gap was not closed (Resing, Xenidou-Dervou, et al., 2012). In Tunteler & Resing's (Tunteler & Resing, 2010) microgenetic study including graduated prompting of geometric analogies the children with a less efficient WMC caught up with their peers with better WMC task performance after training. WMC appears related to training effects in dynamic tests. In the present study we broaden this investigation by examining whether the relationship of the dynamically assessed analogical reasoning skills and WMC extends to affect transfer to other inductive reasoning tasks.

4.1.4 Dynamic measurement of inductive reasoning

The dynamic test of figural analogies we administer contains a pretest, training and posttest. The outcomes of the dynamic test were pretest ability and performance change after training (posttest minus pretest) on the figural analogies task. In addition, we included the children's performance change from pretest to posttest on geometric analogies and seriation transfer tasks, and their ability to solve an analogy construction transfer task administered only on the posttest. We look at change in performance over time, therefore it is important to pay attention to how we measure change because using classical test theory (CTT) scores, such as proportion correct, has received much criticism by psychometricians (e.g., Bereiter, 1963; Embretson, 1991b, 1991a; Lord, 1963; Prieler & Raven, 2002). The main problem with using CTT scores in a dynamic testing context is that when pretest and posttest scores are highly correlated, as is generally the case with repeated measures of the same construct, the change score is unreliable. This of course is unacceptable when one wants to reliably measure change. Furthermore, CTT scores are sensitive to bottom and ceiling effects and the meaning of change scores is dependent upon the examinee's pretest performance. For example, an improvement of four correct solutions may

mean something different on a test of twenty items when the initial score was two or sixteen; a change in scores from two to six may represent greater improvement in understanding than sixteen to twenty. Item response theory (IRT), often referred to as modern test theory, offers solutions for the statistical pitfalls of measuring change with CTT e.g., Embretson & Reise, 2000. IRT scoring in its simplest form, the Rasch model, is based not only on the ability of the person taking the test, but also on the difficulty of the items included in the test. Embretson (1991b) proposed an IRT model, the Multidimensional Rasch Model for Learning and Change (MRMLC), that provides both reliable initial ability and change estimates that can be applied to dynamic testing (e.g., Dörfler, Golke, & Artelt, 2009; Embretson, 1987; Embretson & Prenovorst, 2000) and longitudinal research (e.g., Von Davier et al., 2010). We use this model for estimating the children's pretest abilities and performance change from pretest to posttest.

4.1.5 Current study

In sum, this study investigated the effect of the graduated prompts training method on Rasch-scaled ability and performance change scores of figural analogies and inductive reasoning transfer tasks while examining the role of working memory capacity herein. In accordance with the literature described above we expected (hypothesis 1) the children's performance change on the figural analogies task to be greater in children trained with graduated prompts than when only practicing with the items (see Stevenson et al., submitted 2011a). Transfer of reasoning skills was expected to coincide with initial ability (hypothesis 2a), where transfer effects would be greater in higher ability children (e.g., Tunteler & Resing, 2010). Trained children were expected to show greater transfer effects on the transfer tasks with differed only in content (geometric analogies) or format (analogy-construction) to the trained figural analogies task (hypothesis 2b: (e.g., Roth-Van Der Werf et al.,

2002). Furthermore we expected children with greater WMC to obtain to perform better on the figural analogies and transfer task pretests (hypothesis 3: e.g., Krumm & Buehner, 2008). The final aim was to explore the role of working memory in inductive reasoning transfer.

4.2 Method

4.2.1 Participants

Participants were 64 7-8 year olds (34 girls, 30 boys, M=7.6 years; SD=4.7 months). The children were recruited from three elementary schools located in two midsized towns in the Netherlands. The schools were selected based upon their willingness to participate. All children were native Dutch speakers. Written informed consent was obtained from the parents prior to participation.

4.2.2 Design & Procedure

A pretest-training-posttest control-group design with randomized blocking was employed. Children were blocked into a training or practice group for the AnimaLogica dynamic test based on scores on a visual exclusion test (Bleichrodt et al., 1987) and gender. Children were tested during six weekly sessions.

During the first session, the exclusion task and three working memory tasks were administered. In the next session all children were administered the figural analogies pretest. During the third session two transfer task pretests were administered: geometric analogies and seriation. The fourth session comprised of either training or practice in solving figural analogies. The fifth session consisted of the figural analogies posttest plus an analogy construction transfer task referred to as the reversal task. In the final session the geometric analogies and seriation transfer tasks were re-administered.

AnimaLogica, the dynamic test of figural analogies used in this study, and working memory tasks were administered individually. Classroom-based administration was conducted for the exclusion task and the geometric analogies and seriation transfer tasks.

4.2.3 Instruments

AnimaLogica*: a dynamic test of figural analogical reasoning*

Pretest and Posttest. The figural analogies utilized colored (red, yellow or blue) animal figures, classically presented in 2x2 matrix format (e.g., Stevenson et al., 2009). Drawings of familiar animals occupied three squares and the lower left or right quadrant was empty. The transformations were made on the dimensions: (1) animal, (2) color, (3) size, (4) position, (5) orientation and (6) quantity. The child was asked to choose a picture from five options below to solve the puzzle (A:B::C:D). The five systematically constructed answer options included the correct answer, two partially correct answers (with one incorrect transformation) and two non-analogical answers (with 2 or more incorrect transformations). The test booklets each consisted of 30 items of increasing difficulty.

Training. The training items also consisted of figural analogy matrices. The objects and transformations were the same as the figural analogies task, but instead of multiple-choice items the training items were presented in constructed-response format (see Stevenson et al., under review). None of the 8 training items were identical to the test items. To solve the analogies, the children had to construct the solution from a number of animal cards representing the six transformations; each animal was available in three colors (red, yellow, blue), two possible sizes (large, small) and printed two-sided so by turning the card over the animal's orientation could be changed (looking left or right). Quantity was specified by selecting one or

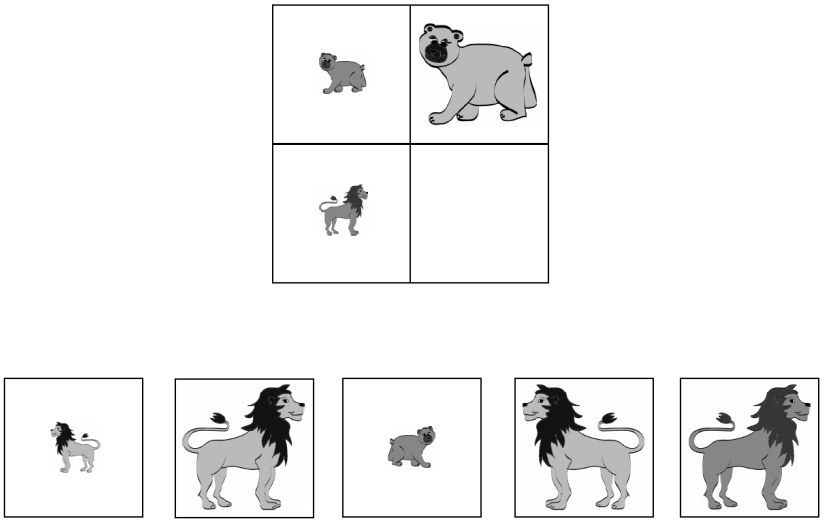

Figure 4.1 *A multiple-choice figural analogy item from* AnimaLogica.

more animal cards and position was selected by placement in the empty box. An example item is shown in Figure 4.1.

Graduated prompting techniques (e.g., Campione & Brown, 1987; Resing, 1993; Resing & Elliott, 2011; Resing et al., 2009) were applied to aid the children in solving the training items. The stepwise instructions 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 (see Stevenson et al., submitted 2011a). A total of five prompts were administered. During the first prompt the child was asked how such an item was solved previously and was provided with more detailed instruction, thereby aiding problem recognition and redefinition. During the second prompt a card was presented which included the general steps to solve the analogies, analogous to Sternberg's (1977) componential analysis: (1) look closely (encoding component), (2)

think about how the animals change (inference component), (3) apply this to solve for the empty box (mapping component) and (4) check your work (justification component). In the third prompt, these components were further emphasized while the examiner worked through the steps on the aid card with the child, explaining with both words and gestures. For example, "What changes from here to here (A:B)?". In the fourth prompt the horizontal and vertical transformations were summarized, emphasizing inference and encouraging mapping. In the final prompt the examiner used scaffolds to help the child systematically solve the problem per transformation, such as "Which animals belong in the empty box?", "Which direction should the dog face?". After each question direct feedback was given, guiding the child step-by-step to the correct solution. 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.

Transfer tasks

The three transfer tasks were selected because each has been used in previous studies assessing inductive reasoning transfer in children. Each task differed from the main task, figural analogies, with regard to content, format and/or measured construct.

Geometric analogies. The geometric analogy task (Hosenfeld & Resing, 1997) consisted of 20 multiple-choice items (see Figure 4.2). The child had to choose the correct answer from five options. The content differs from the figural analogies in that geometric objects instead of animal figures are used. Otherwise the tasks are superficially similar and both require analogical reasoning skills and are presented in multiple-choice format.

Seriation. The seriation task (Durost et al., 1970) consisted of 20 numerical and 14

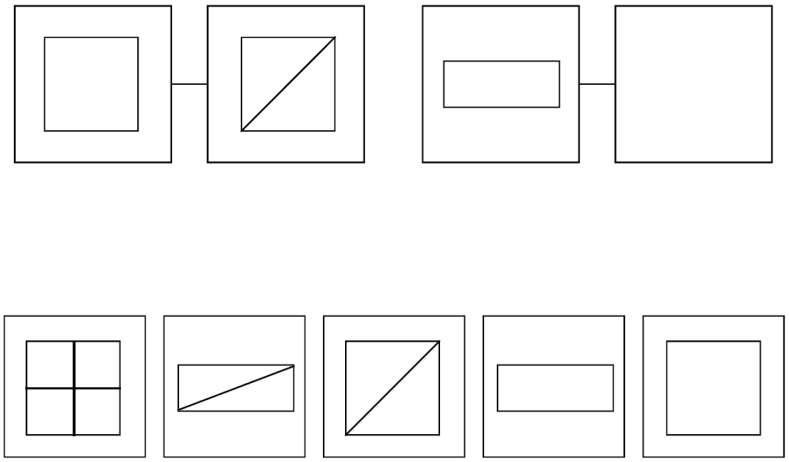

Figure 4.2 *An example item from the geometric analogies transfer task (Hosenfeld et al., 1997).*

geometrical seriation items respectively (see Figure 4.3). The answer to complete the series is selected from four or five options on the right-hand side of the row. The content, geometric objects, is similar to that of the geometric analogies, but different from the animal figures in the figural analogies. This task is also presented in multiple-choice format but requires a different form of inductive reasoning than the figural analogies task, namely series completion, and therefore differs in the measured construct.

Reversal Task. The reversal task is an analogy construction task in which the child is asked to take on the role of teacher (Bosma & Resing, 2006) and construct a matrix analogy for the examiner. The content of this task is the same as the figural analogies task as the same animal figures were used as in the AnimaLogica task, but here the matrix was empty (see Figure 4.4. The format was different because

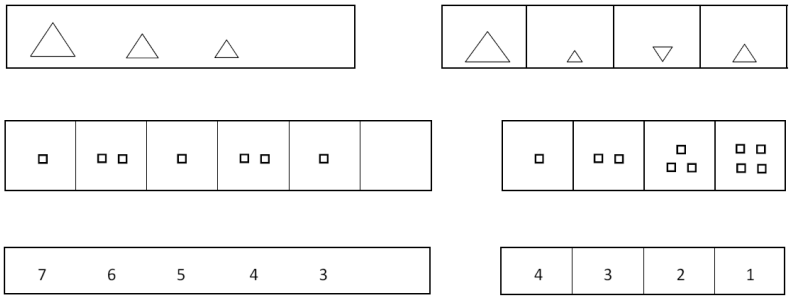

FIGURE 4.3 *Three example items from the seriation transfer task (Durost, et al., 1970).*

the child was asked to construct an analogy and instruct the experimenter on how to solve it. This task requires understanding of analogical reasoning to be able to construct a correct analogy (e.g., Harpaz-Itay et al., 2006).

Working memory

Backward Digit Span. The WISC IV Digit Span backwards (Wechsler, 2003) is considered a measure of verbal working memory capacity (e.g., Süb et al., 2002). The child is asked to repeat a sequence of digits in reverse order.

Listening Recall. The Automated Working Memory Assessment (AWMA, Alloway, 2007) listening recall 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). This task measures verbal working memory.

Spatial Span. In the AWMA (Alloway, 2007) spatial span 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

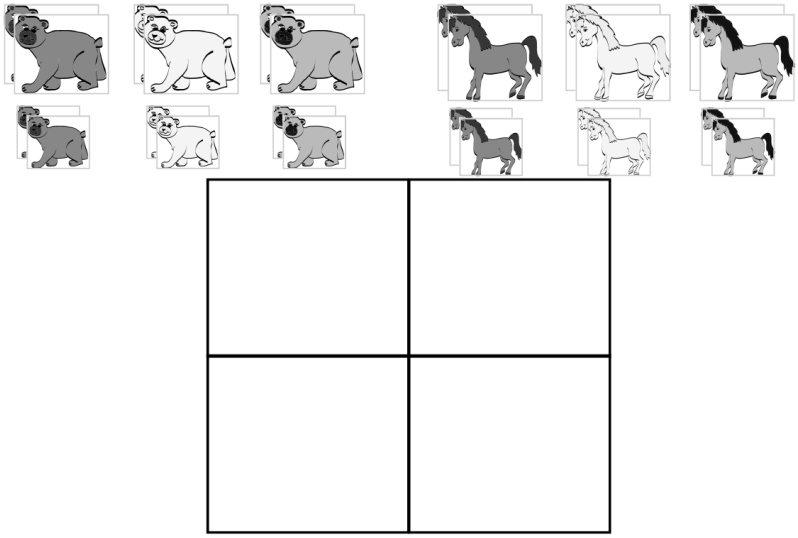

Figure 4.4 *The* AnimaLogica *reversal transfer task (analogy construction).*

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. This task measures visuospatial working memory.

Visual exclusion

Visual exclusion is a subtest of the Revised Amsterdam Children's Intelligence Test (RAKIT, (Bleichrodt et al., 1987) used to measure visual-spatial inductive reasoning ability. The child is shown four abstract geometric figures and asked to choose which one doesn't belong to the other three.

4.2.4 *Scoring*

The children's answers to the figural analogies, geometric analogies and seriation items were based on the selected or constructed answer and scored as correct/incorrect (skipped items were scored as incorrect). Rasch estimates from item response theory (IRT, e.g., Embretson & Reise, 2000) were obtained for the initial ability (pretest performance) and performance change (gain from pretest to posttest) using Embretson's Multidimensional Rasch Model for Learning and Change (MRMLC, Embretson, 1991b, 1991a). Initial analyses were conducted with the `ltm` package for R (Rizopoulos, 2006) and the MRMLC estimates were computed using the `lmer4` package (Bates & Maechler, 2010).

The measure of children's performance on the reversal items was based on a combination of whether they could correctly construct an analogy and the complexity of the analogy, represented by the number of transformations present (e.g., Hosenfeld & Resing, 1997; Mulholland et al., 1980; Stevenson et al., 2009, 2011). The resulting score on this analogy construction task was correctness (1/0) x number of represented transformations (1-6).

4.3 Results

Before conducting analyses to answer the research questions we first describe the psychometric properties of the Rasch-scaled tests and items. Furthermore, we check whether the children in the two conditions differed in cognitive functioning or age prior to testing.

4.3.1 *Psychometric Properties*

Pretests and posttests were administered for the figural analogies (FA), geometric analogies (GA) and seriation (SR) tasks. Cronbach's measure of internal consistency

on the pretests was $\alpha = .81$, $\alpha = .92$ and $\alpha = .91$ for FA, GA and SR tasks respectively. For the posttests this was $\alpha = .83$, $\alpha = .92$ and $\alpha = .91$ for the FA, GA and SR tasks.

Before applying the MRMLC model, first the independent Rasch model parameters were estimated for the pretests and posttests using marginal maximum likelihood (MML) estimation. The parametric Bootstrap goodness-of-fit test in the `ltm` package was used to investigate model fit. The Pearson's χ^2 statistic (based on a comparison with 50 generated datasets) indicated that the Rasch model fit of the pretests and posttests were acceptable ($p > .05$) with the exception of the seriation pretest ($p = .02$). The correlations between the item parameters of the pretests and posttests were very strong for each of the tasks, $r_{FA} = .76$, $r_{GA} = .87$ and $r_{SR} = .91$, therefore we considered the application of Embretson's MRMLC model appropriate. The range of the MRMLC item difficulty parameters was -2.60 to 3.40 ($M = .83, SD = 1.60$) for the FA task, -3.03 to 1.33 ($M = -.20, SD = 1.02$) for GA task and -4.42 to 1.69 ($M = -.75, SD = 1.18$) for the SR task.

4.3.2 *Initial Group Comparisons*

The children's average age ($F(1, 62) = 2.13, p = .15$), initial level of inductive reasoning (visual exclusion: $F(1, 62) = .27, p = .61$), working memory capacity (backward digit span (BDS): $F(1, 62) = .23, p = .64$; listening recall (LR): $F(1, 62) = .02, p = .88$; spatial span (SS): $F(1, 62) = .48, p = .49$) and pretests (figural analogies: $F(1, 62) = .36, p = .55$; seriation: $F(1, 62) = .00, p = .98$) did not differ between conditions (see Table 4.1 for basic statistics). Initial performance on geometric analogies pretest differed significantly between conditions: $F(1, 62) = 5.45, p = .02$.

4.3.3 *Effect of graduated prompting on figural analogy solving*

Our first research question concerned the effect of the graduated prompts training in improving the children's performance on the figural analogies task. We expected

TABLE 4.1 *Basic statistics of age, exclusion, working memory and pretest scores (*MRMLC *ability estimates) of figural analogies, geometric analogies and seriation per condition.*

	Training (N=32)		Practice (N=32)		Total (N=64)	
	M	SD	M	SD	M	SD
Age	92.38	5.12	90.66	4.27	91.52	4.75
Visual exclusion	16.06	6.92	16.97	7.16	16.52	7.00
Working memory:						
Digit span backwards	6.71	1.90	6.94	1.83	6.82	1.85
Listening recall	12.58	3.75	12.71	2.76	12.65	3.27
Spatial span	16.48	6.41	17.55	5.73	17.02	6.06
Pretest score:						
Figural analogies	.0772	.7717	-.0591	1.0299	.0091	.9054
Geometric analogies	-.4604	1.4767	.4415	1.6109	-.0094	1.5990
Seriation	-.0203	1.2422	-.0118	1.2395	-.0161	1.2310

(1) that graduated prompts techniques would lead to greater improvement in analogical reasoning scores than practice alone. This was investigated using an analysis of variance (ANOVA) with figural analogy performance change estimates as the dependent variable and pretest score as the covariate. There was a between-subjects effect for condition ($F(1, 61) = 3.99, p = .05, \eta_p^2 = .06$) indicating that the conditions differed in their degree of improvement. The covariate, pretest score, did not affect the change score: $F(1, 61) = 1.14, p = .29, \eta_p^2 = .02$. Inspection of the means and standard deviations (see Table 4.3.3) shows that the children in the training condition obtained significantly higher performance change scores than those in the practice condition, confirming hypothesis 1.

Table 4.2 *Basic statistics of performance change from pretest to posttest (*MRMLC*) and reversal task performance.*

	Training (N=32)		Practice (N=32)		Total (N=64)	
	M	SD	M	SD	M	SD
Change scores:						
Figural analogies	.0673	.2683	-.0808	.3444	-.0067	.3152
Geometric analogies	-.0199	.4995	-.0192	.5640	-.1910	.6873
Seriation	-.0081	.8012	.0007	.4299	-.0037	.4623
Reversal task score:						
Analogy Construction	2.47	2.19	1.69	1.51	2.08	1.91

4.3.4 Effect of graduated prompting on transfer

The second research question related to the children's ability to transfer learned figural analogical reasoning skills to geometric analogies (GA), seriation (SR) and an analogy construction (AC) task. We expected (2a) transfer effects to be related to the children's pretest scores and expected (2b) to find a training effect on the transfer tasks of similar content or format (GA and AC). For the GA and SR tasks transfer was ascertained using the change score – i.e. degree of improvement from pretest to posttest. For the AC task, which was not pretested, the reversal task score (see 4.2.4 Scoring) was used as a transfer measure.

Before testing our hypotheses we computed the correlations between the FA pretest scores and performance on the three transfer tasks. The FA pretest score correlated strongly with the pretest scores on the GA ($r = .57, p < .001$) and SR tasks ($r = .63, p < .001$). The AC score was moderately correlated with the FA pretest score ($r = .37, p = .003$).

To investigate the relationship of transfer with condition and FA pretest performance a MANCOVA (3 transfer tasks x 2 conditions) with GA change

estimates, SR change estimates and AC scores as dependent variables with figural analogy pretest score as covariate was conducted. An effect was found for FA pretest score on AC performance ($F(1,61) = 9.33, p = .003, \eta_p^2 = .13$), but not on the change scores of GA ($F(1,61) = .11, p = .74, \eta_p^2 = .00$) or SR ($F(1,61) = 1.71, p = .20, \eta_p^2 = .03$). Transfer effects appear only partially related to FA pretest scores (Wilks' $\lambda = .85, F(3,59) = 3.44, p = .02, \eta_p^2 = .15$); hypothesis 2a is only partly accepted. Results show that condition does not lead to a differential effect on transfer (Wilks' $\lambda = .96, F(3,59) = .79, p = .51, \eta_p^2 = .04$), hypothesis 2b is rejected.

4.3.5 Role of working memory in analogical reasoning ability and transfer

Our third research question pertains to the role of working memory in analogical reasoning ability and transfer. We expected (3a) working memory capacity to be related to the children's performance on all tasks. We also explored whether (3b) WMC was positively related to transfer effects.

First correlations were used to examine the relation between WMC and children's performance and change on all tasks. Backward digit span (BDS) showed a moderate correlation with FA, GA and SR (see Table 4.3.5), confirming hypothesis 3a. The correlations of the change estimates and AC scores with WMC were not significant; therefore hypothesis 3b could be rejected.

The three working memory tasks were not strongly correlated (BDS, LR $r = .10, p = .46$; BDS, SS $r = .34, p = .01$ and LR, SS $r = .34, p = .07$). In order to gain greater insight into the WM components involved in each of our experimental tasks, we further investigated whether combinations of the working memory measures explained significantly greater variance in pretest scores than just BDS. Hierarchical regression analyses were conducted with BDS entered as the first predictor and LR or SS as the second variable. In the case of figural analogies the best fitting model included listening recall ($\Delta R^2 = .061$) in addition to BDS explaining 16.6% of the

Table 4.3 *Correlations of working memory measures and pretest and change scores of figural analogies, geometric analogies, seriation and reversal analogy construction score.*

	Backward digit span	Listening Recall	Spatial Span
Pretest score:			
Figural analogies	.323*	.277*	.268*
Geometric analogies	.347*	.101	.303*
Seriation	.385*	.206	.345*
Change score:			
Figural analogies	-.003	-.081	.145
Geometric analogies	.063	-.168	.044
Seriation	.113	.012	.069
Reversal task score:			
Analogy Construction	.104	.215	-.041

*$p < .05$

variance (see Table 4.3.5). For GA and SR, neither LR or SS explained significant additional variance, although in both cases when BDS was excluded from the analyses SS was the best predictor (see Table 4.3.5).

4.4 Discussion

The main aim of this study was to investigate the learning and transfer of analogical reasoning skills in a dynamic testing context and explore the role of working memory capacity herein. We compared the learning and transfer of inductive reasoning skills of children who were trained during dynamic testing with graduated prompts or practiced without feedback on a figural analogies task. As with previous studies (e.g., Siegler & Svetina, 2002; Tunteler & Resing, 2010), we found that trained children showed greater progression in analogy solving than children in the practice condition. Furthermore, performance on the figural analogy matrices

TABLE 4.4 *Results of hierarchal linear regression analyses predicting pretest scores from working memory measures.*

Dependent variable		Predictor	B	SE B	β
Figural analogies					
	Step 1	Backward digit span	.148	.059	.300*
	Step 2	Listening recall	.070	.033	.249*
Geometric analogies					
	Model 1	Backward digit span	.299	.104	.347**
	Model 2	Spatial span	.071	.025	.345**
Seriation					
	Model 1	Backward digit span	.259	.080	.385**
	Model 2	Spatial span	.080	.032	.303*

*p < .05; **p < .01

pretest was strongly related to performance on each of the transfer tasks: geometric analogies, seriation and analogy construction. This coincides with previous research as the relationship between these tasks has been emphasized in numerous studies (e.g., Carpenter et al., 1990; Roth-Van Der Werf et al., 2002; Sternberg & Gardner, 1983). Transfer of analogical reasoning skills to the reversal situation in which the child constructed an analogy for the examiner was related to initial ability on the figural analogies tasks, where more complex analogies were constructed by the children with higher pretest scores. The findings on the reversal task were in line with Siegler's theory (Siegler, 2006) that greater mastery of task strategies increases the chances of knowledge transfer to a novel situation in children. Yet as with previous research on the effect of the short graduated prompts training on transfer of inductive reasoning skills (e.g., Tunteler & Resing, 2010; Roth-Van Der Werf et al., 2002), we found that children in the training condition showed a similar degree of improvement from pretest to posttest on transfer tasks with dissimilar content

as the children who only practiced with the items. A possible explanation for our results and those of previous studies where training on a different task does not affect transfer of knowledge to similar tasks stems from Opfer & Thompson's (2008) practice interference hypothesis. Their theory suggests that practice using incorrect solution strategies, which often occurs during pretesting, which was included in the present and previous studies, impedes transfer. This could explain why transfer effects were only found on the reversal task which was not pretested, but not on the other two transfer tasks in this study. In the assessment of transfer within dynamic tests, which often comprise a pretest-training-posttest format, it is perhaps advisable not to pretest the transfer tasks. Instead a selection of transfer tasks that measure similar skills to the tested task may provide more reliable measures. The effect of initial ability could be accounted for using the pretest scores of the trained task, which indeed correlated with performance on the analogy construction (reversal) task in the present study. However, greater transfer of knowledge has been demonstrated in other research in which pretests were included but more extensive training was provided (e.g., Harpaz-Itay et al., 2006; Klauer & Phye, 2008; Rittle-Johnson, 2006; Siegler & Svetina, 2002). We therefore advise including more training sessions to investigate whether trained children would show greater transfer on a group level than practice or control groups to further verify the effects of the graduated prompting procedure.

The goal of dynamic testing, however, is to ascertain the amount of learning and transfer an individual can achieve after a short training procedure in order to gain insight into learning efficiency. In order to assess this we used item response theory (IRT) Rasch estimates of the degree of performance change the children showed from pretest to posttest. These estimates provide a more accurate picture of proficiency change by avoiding statistical pitfalls of traditional scores, such as percentage correct, where change scores are unreliable and bottom or ceiling effects

could warp the degree of performance change (e.g., Bereiter, 1963; Prieler & Raven, 2002). We used Embretson's (1991b, 1991a) MRMLC model which provides reliable change estimates to measure training and transfer effects on the pretested tasks. Our results show great variability in initial ability and performance change on of each of the inductive reasoning tasks and we therefore investigated whether working memory capacity could be a source of individual differences.

A great deal of research with adults has demonstrated that working memory capacity is strongly related to fluid intelligence and inductive reasoning (e.g., Ackerman, Beier, & Boyle, 2005). It is also postulated to be a bottleneck in children's analogical reasoning (Richland et al., 2006; Thibaut, French, & Vezneva, 2010). Our results coincide with this as we found moderate correlations between working memory measures and the children's initial ability levels on all three inductive reasoning tasks. This relationship ($r \approx .35$) was not as strong as in adult populations, but similar to that found in other research with children (Alloway, Gathercole, Willis, & Adams, 2004; Hornung, Brunner, Rueter, & Martin, 2011; Tillman, Nyberg, & Bohlin, 2008). Verbal WM played a stronger role in the solution of figural analogies and visuo-spatial WM contributed more to performance on the geometric analogies and the geometric and numerical seriation task. These findings are in line with Hornung et al. (2011) where substantial relationships were found between the verbal and visuospatial WM factors with young children's performance on Raven's colored matrices – a task which among other traits also requires inductive reasoning to solve. However, given their conclusion that short-term memory best explains the relationship between working memory and Raven performance, it is advisable to include short-term memory in future investigations of the role of memory in children's performance on inductive reasoning tasks.

From the literature and our results we can conclude that WMC is related - to a certain degree - to inductive reasoning ability in children (Engel de Abreu, Conway,

& Gathercole, 2010; Hornung et al., 2011; Richland et al., 2006; Tillman et al., 2008; Tunteler et al., 2008; Tunteler & Resing, 2010). Given the importance placed upon WMC in cognitive and psychoeducational assessment (Hatcher, Snowling, & Griffiths, 2002; Martinussen, Hayden, Hogg-Johnson, & Tannock, 2005; Pickering & Gathercole, 2004) the question arises whether WMC can explain individual differences in the amount of learning and transfer a child demonstrates in a dynamic assessment procedure. In this study, we found WMC was unrelated to the children's improvement on the trained task or degree of transfer to related tasks after training. It appears that WMC does not sufficiently explain individual differences in learning or transfer in a dynamic testing context. Our analysis of the role of WMC was exploratory and the study comprised of a small sample, therefore more extensive research is needed to substantiate our findings.

Inductive reasoning ability and WMC are well-established constructs in cognitive ability tests and known to be related. Performance change and ability to transfer knowledge to novel situations, such as in the reversal task, are less often included in the assessment of intellectual abilities (Bosma & Resing, 2006; Elliott et al., 2010). Our finding that change scores and knowledge transfer are only somewhat related to initial ability and unrelated to WMC indicates that these may be separate constructs and important in the assessment of learning and cognitive potential. Further research should focus on the relevance of change scores and performance on transfer tasks in psychoeducational assessment – whether these constructs provide a better picture of a child's capabilities and potential.

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.

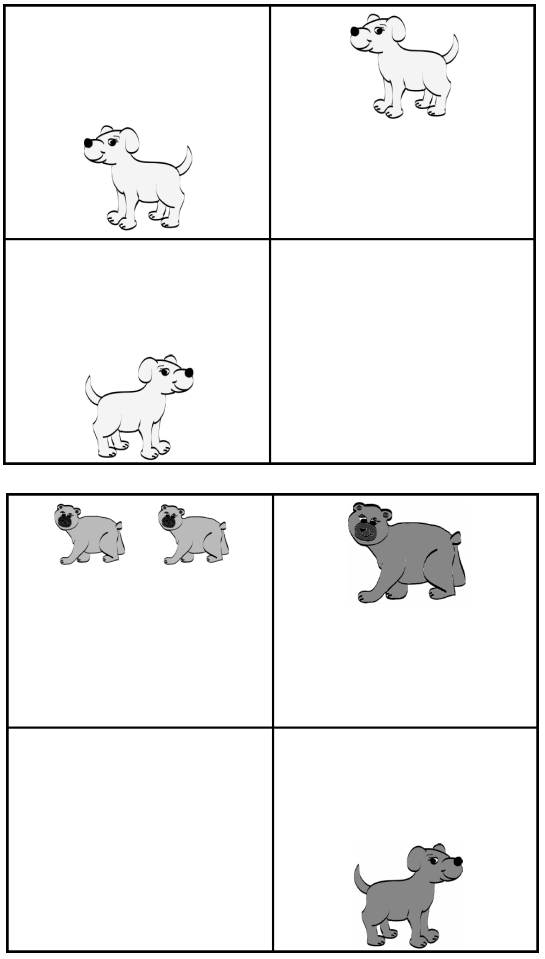

FIGURE 5.1 *Examples of figural matrix analogies used in* ANIMALOGICA. *Top figure contains two transformations (horizontal: position; vertical: orientation). Bottom figure contains six transformations (horizontal: color, quantity and size; vertical: animal, orientation and position).*

*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 26^{th} and 50^{th} percentile whereas 'C' scores (sufficient) indicate 51^{st} to 75^{th} percentile performance. 'D' (weak) and 'E' (very weak) scores fall within the lowest 25% – 'D' scores indicate performance with the 11^{th} to 25^{th} 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.

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

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;

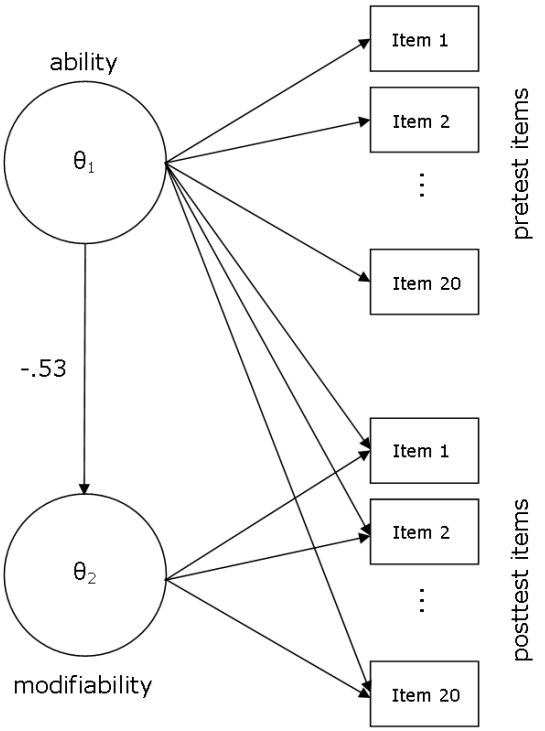

FIGURE 5.2 *Structural equation model of the relationship between ability and modifiability from* MRMLC *(Embretson, 1991a) applied in Model M1b.*

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.

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

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.

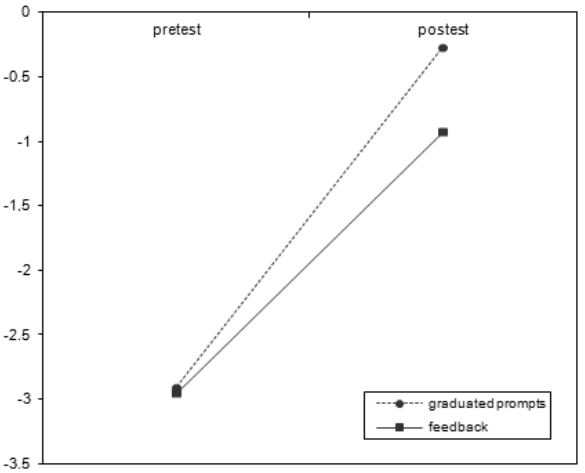

FIGURE 5.3 *Plot of person logits on an average item (four transformations) for both training conditions from pretest to posttest (M2b).*

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.

CHAPTER 6

Dynamic measures of analogical reasoning predict children's math and reading achievement

This chapter is based on Stevenson, C. E., Heiser, W. J. & Resing, W. C. M. (under review). Dynamic measures of analogical reasoning predict children's math and reading achievement.

Abstract

Dynamic testing is an assessment approach that aims to gauge cognitive potential by incorporating training into the testing process. The purpose of this study was to investigate the predictive power of dynamic outcomes compared to traditionally administered (i.e. static) measures of analogical reasoning on children's achievement in reading and math. 253 first and second graders (M=6.99; SD=.73 years) were administered a dynamic test of analogical reasoning comprising a pretest-training-posttest design. Performance on standardized national scholastic tests, categorized from A (very good) to E (very weak), of reading and math were gathered at three time points within one year of dynamic testing. A random-intercepts model for ordinal longitudinal data indicated that the dynamic measure of performance change from pretest to posttest was a better predictor of achievement in math and reading than the static pretest measure. Dynamic measures may prove useful in for educational psychologists when assessing learning ability and cognitive potential.

Acknowledgments

We thank Femke Stad and Carlijn Bergwerff for their assistance in data collection, coding and DT training. We also thank Anna Heethuis and HCO (The Hague Center of Educational Advice) – especially Roel Verdel – for their contribution in data collection. Special thanks to Hailemichael M. Worku for his help selecting the appropriate statistical model for longitudinal ordinal data.

6.1 Introduction

Dynamic testing is considered a diagnostic method that focuses on cognitive potential rather than previous learning (Elliott, 2003; Sternberg & Grigorenko, 2002). Dynamic assessment diverges from static testing, of which an IQ test is a typical example, in that feedback is provided by the examiner during testing in order to facilitate learning and gain insight into learning efficiency and instructional-needs (Elliott et al., 2010). Dynamic testing often comprises a pretest-training-posttest design and can provide indices to examine an individual's potential for learning such as (1) performance change following training (e.g., Day et al., 1997; Pena et al., 2001; Stevenson et al., submitted 2011a; Tzuriel, 2001), (2) the amount of instruction required to solve training tasks (e.g., Bosma & Resing, 2012; Campione & Brown, 1987; Resing & Elliott, 2011), and (3) the ability to spontaneously transfer these newly developed skills to other problems (Ferrara et al., 1986; Resing, 1993, 1997; Stevenson, Hickendorff, Heiser, Resing, & De Boeck, submitted 2012a). Previous research indicates that individual differences in performance on dynamic measures may provide additional information on children's present and future attainment at school (for a review see Caffrey et al., 2008). Yet, further evidence demonstrating the additional value of dynamic testing is necessary to enable more wide-spread acceptance in psycho-educational assessment (Beckmann, 2006; Caffrey et al., 2008; Grigorenko & Sternberg, 1998). In the present study we investigated the predictive power of these three dynamic measures obtained from a dynamic test of analogical reasoning on children's achievement of reading and math at school.

Children are often tested from early on in their school careers given that cognitive assessment scores are considered to be good predictive measures of school achievement (Balboni et al., 2010; Sternberg et al., 2001). Furthermore, school psychologists often use scores on these conventional, static tests in an attempt to identify cognitive weaknesses so that these can be remediated with timely

intervention (Caffrey et al., 2008; Resing, 1997). Yet, critics argue that a pattern of weakness or great potential despite low scores may go undetected given the static nature of these tests (Fabio, 2005; Jeltova et al., 2007; Grigorenko, 2009a; Haywood & Lidz, 2007). Researchers and practitioners have introduced dynamic testing, where the rate and process of learning are emphasized, in order to remedy these shortcomings (Carlson & Wiedl, 1992; Elliott, 2003; Sternberg & Grigorenko, 2002). There is some evidence that dynamic tests provide additional useful information pertaining to an individual's cognitive potential and instructional-needs (e.g., Bosma et al., submitted; D. F. Fuchs, Compton, Fuchs, Bouten, & Caffrey, 2011; Jeltova et al., 2011; Resing, Xenidou-Dervou, et al., 2012). For example, dynamic measures may provide additional predictive value of school achievement in reading (e.g., Bynre, Fielding-Barnsley, & Ashley, 2000; D. F. Fuchs et al., 2011; Resing, 1993; Swanson, 2011b), math (e.g., Beckmann, 2006; Jeltova et al., 2011; Meijer, 1993; Resing, 1993; Sittner Bridges & Catts, 2011) and other school achievement topics such as geography (Hessels, 2009). However, other studies do not consistently show advantages of dynamic measures in predicting school achievement (e.g., Coventry, Byrne, Olsen, Corley, & Samuelsson, 2011; Speece, Cooper, & Kibler, 1990; Swanson, 1994; Thatcher-Kantor, Wagner, Torgensen, & Rashotte, 2011). Caffrey et al. (2008) concluded that the predictive value of dynamic measures was greatest in populations of students with learning disabilities and when criterion-referenced rather than norm-referenced tests were used as the dependent variable. In the present study we aimed to extend these findings by comparing the predictive value of dynamic measures of analogical reasoning on norm-referenced school achievement in typically developing children.

In the studies on the predictive value of dynamic testing a variety of measures have been used in attempts to predict achievement. For example, Swanson's (1994, 2011b) predictors included gain scores, Resing's (1993) predictors included

instructional-needs during training and Rutland & Campbell's (Rutland & Campbell, 1995) also analyzed the predictive value of transfer. It is unclear which dynamic measure is most useful. The instructional-needs measure used by Resing (Resing, 1993) was also included in this study. This was derived from the graduated prompts training procedure in which increasingly elaborate instructions are given when the child is unable to solve the problem independently (e.g., Bosma & Resing, 2012; Resing & Elliott, 2011). The number of prompts required provides an indication of how much instruction a child needs to reach a particular performance level and has demonstrated additional predictive value to static measures on reading and math achievement (Resing, 1993). Transfer, the ability to spontaneously generalize a problem-solving approach taught in one context (such as during training) to a different but applicable situation (Barnett & Ceci, 2002), is measured in the present study using an analogy construction task. Here the child takes on the role of the teacher and "teaches" the examiner how to solve their analogy (Bosma & Resing, 2006). Such tasks seem to measure depth of understanding and divergent thinking (Jaarsveld, Lachmann, & Van Leeuwen, 2012) and may therefore be related to school performance (Vock, Prekel, & Holling, 2011). In the current study we compare the predictive value of three dynamic measures (instructional-needs, transfer and gain) and one static measure in the prediction of young children's achievement in reading and math.

In dynamic testing it is necessary to be careful of how to measure change because classical test theory (CTT) gain scores, such as proportion correct, have received much criticism by psychometricians (e.g., Bereiter, 1963; Embretson, 1991b, 1991a; Lord, 1963; Prieler & Raven, 2002). The main problem with using CTT scores in a dynamic testing context is that when pretest and posttest scores are highly correlated, as is generally the case with repeated measures of the same construct, the change score is unreliable. Furthermore, the meaning of change

scores is dependent upon the testee's pretest performance. Item response theory (IRT) offers solutions for the statistical pitfalls of classical ways of measuring change (e.g., Embretson & Reise, 2000). Embretson (Embretson, 1991b) proposed an IRT model, the Multidimensional Rasch Model for Learning and Change (MRMLC), that provides both reliable initial ability and change estimates that can be applied to dynamic testing (e.g., Dörfler et al., 2009; Embretson, 1987; Embretson & Prenovorst, 2000) and longitudinal research (e.g., Von Davier et al., 2010). In the current study we used this model to estimate the children's pretest ability and performance change from pretest to posttest.

Dynamic tests often include fluid reasoning tasks (e.g., Ferrara et al., 1986; Resing & Elliott, 2011) because of their association with general intelligence, and because they are considered to assess the capacity to solve new problems based on learning how to find rules when solving previous more familiar problems (Carpenter et al., 1990; Primi, 2001). Fluid reasoning ability has been shown to be a good predictor of school achievement for both reading (Stanovich, Cunningham, & Freeman, 1984; Ferrer et al., 2007) and math (Primi et al., 2010; Taub et al., 2008). In the current study we administered figural analogy matrices, a type of fluid reasoning task (e.g., Freund & Holling, 2011b), to 6-8 year-old children and investigate whether children's static or dynamic test results best coincided with their performance on national assessments of reading and math.

In the current study, first and second grade children were dynamically tested on a figural analogies task. Their performance on the Netherlands national school assessments of reading and math were collected at three time points within one year of dynamic testing, with six months between each assessment. We tested the hypotheses that dynamic test measures, in the form of (a) performance change, (b) training and (c) transfer task performances better predict the children's school achievement than a static (i.e. pretest score) measure of figural analogical reasoning.

6.2 Method

6.2.1 Participants

Participants in this study were 253 6-8 year olds (140 girls, 113 boys; M = 83.92, SD = 8.82 months), who were dynamically tested at Time 1. They came from 18 schools in urbanized cities in the Western parts of the Netherlands. At Times 2 and 3 school achievement data was available for 182 and 141 children respectively. None of the children were diagnosed with learning disabilities or behavior problems prior to dynamic testing. Written informed consent was obtained from the parents.

6.2.2 Instruments

AnimaLogica*: a dynamic test of figural analogical reasoning*

Pretest and Posttest. The figural analogies utilized colored animal figures, classically presented in 2x2 matrix format on a computer screen (see Figure 6.1). The animals occupied three squares and the lower left or right quadrant was empty. Transformations were made on the dimensions: (1) animal, (2) color, (3) size, (4) position, (5) orientation and (6) quantity. The children had to construct the solution using a computer mouse to drag & drop animal figures representing the six transformations into the empty box. 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.

The test booklets consisted of 20 items of varied difficulty. The pretest and posttest contained item isomorphs – comprising the same transformations and difficulty, but different animals and colors. Cronbach's measure of internal consistency for the pretest and posttest were α = .83 and α = .90 (N = 514).

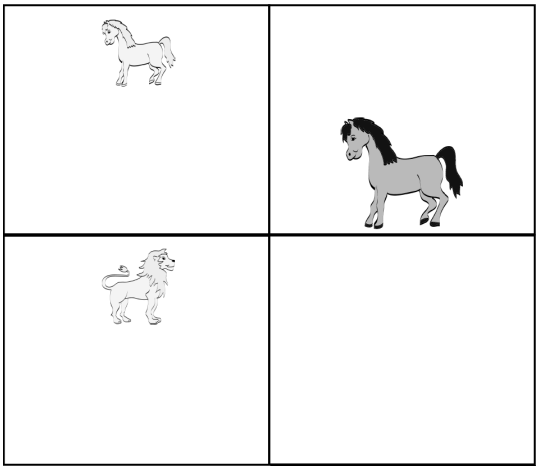

Figure 6.1 *Example item from* AnimaLogica.

With regard to construct validity, the pretest correlates highly with other cognitive measures: Raven Standard Progressive Matrices (Raven et al., 2004), $r = .60$ ($N = 253$) and Automated Working Memory Assessment (awma, Alloway, 2007) listening recall, $r = .42$ ($N = 252$) and spatial span, $r = .45$ ($N = 252$).

Training. The training consisted of the same figural analogy matrices. The 10 training items did not occur in the tests. During the training phase the children received instruction in analogy solving according to the graduated prompts method (e.g., Campione & Brown, 1987; Resing, 1993; Resing & Elliott, 2011; Resing et al., 2009; Stevenson et al., submitted 2011a). The stepwise instructions 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 (see Stevenson et al., submitted 2011a. 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. The reliability of the training items scale was $\alpha = .84$ ($N = 379$).

Transfer. The transfer task was an analogy construction task presented in reversal format in which the child was asked to take on the role of teacher (Bosma & Resing, 2006) and construct a matrix analogy and explain how to solve it. The same animal figures were used as in the test and training sessions, but here the matrix was empty (see Figure 6.2).

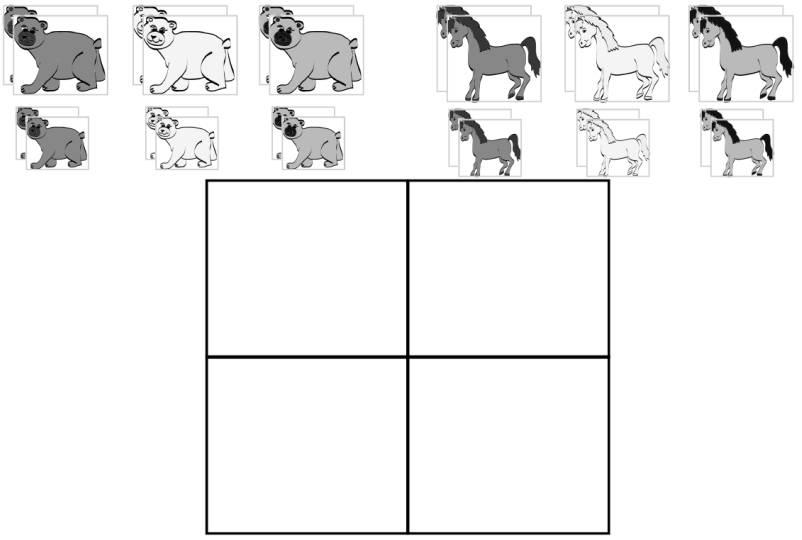

FIGURE 6.2 *The reversal transfer task (analogy construction) from* ANIMALOGICA.

Scoring. ANIMALOGICA provides four scores: (1) pretest, (2) performance change, (3) training and (4) transfer. The pretest score is considered a static test score, whereas performance change is a dynamic score that quantifies the difference in performance from pretest to posttest (posttest minus pretest). The correct/incorrect construction

of the figural analogies on the pretest and posttest were used to compute the initial ability and performance change scores on a item response theory scale (IRT, Embretson & Reise, 2000) using Embretson's Multidimensional Rasch Model for Learning and Change (MRMLC, Embretson, 1991b. The MRMLC estimates were computed for the entire dataset (N=514) using the `lmer4` package (Bates & Maechler, 2010) for R (e.g., Stevenson et al., submitted 2012a).

The training score is a dynamic score and quantifies the amount of help (max. 5 prompts per item) required by the child to solve the training items (e.g., Ferrara et al., 1986; Resing, 1993, 1997). The transfer score is the third dynamic score and quantifies the child's performance on the reversal task. This is based on a combination of whether the child could correctly construct an analogy and the complexity of the analogy, represented by the number of transformations present (e.g., Hosenfeld & Resing, 1997; Mulholland et al., 1980; Stevenson et al., 2009, 2011). The resulting score on this analogy construction task was correctness (1/0) x number of represented transformations (1-6) (e.g., Stevenson, Heiser, & Resing, submitted 2011b).

Standardized scholastic achievement tests

The children each took part in biannual scholastic achievement assessments administered in January and June of each school year (CITO , 2010b, 2010d). 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 scores are based on national norms and range from A to E. An 'A' is categorized as a very good, indicating performance falls 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%. Schools are allowed to choose which tests to administer therefore, for some children at certain time points only math scores and in other cases only reading scores were available.

6.2.3 Design & Procedure

Each child was tested individually in a quiet room at the school for the four weekly sessions of the dynamic test. Each session lasted approximately 20 minutes and total testing time comprised less than 1.5 hours. The pretest was administered during the first session. The training procedure was administered during the second and third sessions. The posttest and the transfer task were administered during the fourth (final) session. Trained graduate students and educational psychologists administered the dynamic test.

The scholastic achievement tests were administered by the child's teacher in the classroom. The first measure took place 3 weeks prior to dynamic testing. The second and third measurements were administered six months and one year later.

6.2.4 Statistical Model

A random intercepts model for repeated ordinal data was used to analyze whether static or dynamic variables best predicted the children's school achievement during one year for reading and math. This form of probabilistic odds model was chosen because the scale of the dependent variable, the standardized test scores in each of the school subjects, consisted of five ordered categories ranging from 'E' lowest performance to 'A' highest performance. Furthermore, this model can account for the longitudinal nature of our design and deals adequately with the missing values inherent to our data (Molenberghs & Verbeke, 2005).

Math and reading achievement were modeled separately. Let Y_{st} denote the ordinal achievement category ('A'=5,'B'=4,'C'=3,'D'=2,'E'=1) at time t (t = 0: at

time of dynamic testing, $t = 1$: six months later, $t = 2$: one year later) for child s. The random-intercept model of the cumulative probability of $Y \leq i$ ($i = 1, 2, 3, 4, 5$) with K predictors and their interactions with continuous time t is defined by equation (1).

$$log\left[\frac{P(Y_{st} \leq i)}{1 - P(Y_{st} \leq i}\right] = u_s + \beta_0 t + \sum_{k=1}^{K} \beta_k X_{stk} \tag{6.1}$$

The random effect, u_s, is child-specific and assumed to be normally distributed ($u_s \sim N(0, \sigma^2)$) and the same for each cumulative probability including a nonnegative correlation between the observations of a particular child. β_0 denotes the main effect of time. K predictor variables are: (1) pretest score, (2) change score, (3) training score or (4) transfer score for which both main effects and interaction effects with time are included. The initial score (predictor 1) represents static test performance and predictors 2-4 represent dynamic test performance. According to our main hypothesis, that dynamic test results provide a better indication of a child's achievement category than static measures, we expected that β_2, β_3 or β_4 – associated with the three dynamic scores – would better predict $Y \leq i$ than β_1, the static predictor.

6.3 Results

6.3.1 Descriptive Statistics

The descriptive statistics of each of the variables are presented in Table 6.1. The large standard deviations for each of the AnimaLogica scores indicate that individual differences are present in performance on each of these measures. The reading and math achievement scores for each time point on average fall in between categories 3 ('C') and 4 ('B'). This is as expected as the 50th percentile rank is at the border between these two categories. The sample means are slightly higher than 3.5 which

TABLE 6.1 *Descriptive statistics of static and dynamic test scores (predictor variables) and reading and math achievement (dependent variables).*

	N	Min.	Max.	Median	M	SD	Skewness	Kurtosis
AnimaLogica								
Static pretest	253	-2.96	5.00	.05	.21	1.57	0.49	-0.03
Dynamic change	253	-2.77	3.25	.22	.32	1.03	0.27	-0.08
Dynamic training	253	0	50	15	18.20	12.91	0.65	-0.58
Dynamic transfer	253	0	9	2	2.21	2.12	0.51	-0.71
Reading Achievement								
Reading Time 1	211	1	5	4	3.67	1.29	-0.59	-0.76
Reading Time 2	167	1	5	4	3.70	1.24	-0.62	-0.54
Reading Time 3	141	1	5	4	3.75	1.07	-0.52	-0.55
Math Achievement								
Math Time 1	237	1	5	4	3.58	1.30	-0.51	-0.87
Math Time 2	182	1	5	4	3.78	1.31	-0.73	-0.63
Math Time 3	138	1	5	4	3.65	1.32	-0.49	-1.00

indicates that on average the participants in the present sample perform slightly better than the national average.

The correlations between each of the measures are presented in Table 6.2. Here we see that the AnimaLogica measures are generally weakly, but significantly inter-correlated except for dynamic change and transfer which are not correlated. Another exception is the very strong correlation between the static pretest and the training score. The correlation between reading achievement at times 1 and 2 is strong, and from time 2 to 3 moderate. Math achievement scores from time 1 to 2 to 3 each are strongly and positively correlated indicating that children either remained in the same category or generally changed categories in the same direction. Correlations between math and reading categories are moderate to strong when measured at the same time point and negligible to weak when comparing different time points. Reading achievement and AnimaLogica dynamic change have a positive but weak association for times 1 and 2. A weak to marginal negative association is present between reading achievement and the training score. The correlations between math achievement and AnimaLogica scores are strongest for the training scores, but also moderate to strong for the pretest scores. Furthermore the dynamic change score has a positive weak association with math achievement at time points 1 and 2.

6.3.2 Statistical Modeling

Reading achievement

Table 6.3 presents the Maximum Likelihood fit results of the proportional odds model from Equation (1) for reading achievement over time computed with the SAS NLMIXED procedure as described by Molenberghs & Verbeke (2005) comprising 516 observations (from 3 occasions for 232 students with 243 missing values). The model fit statistics were AIC = 840, BIC = 875, $-2LogLikelihood$ = 820 with 10

TABLE 6.2 *Correlations between reading and math achievement (dependent variables) and static and dynamic test scores (predictor variables).*

		1	2	3	4	5	6	7	8	9	10
AnimaLogica											
1. Static pretest	r	1									
	N	253									
2. Dynamic change	r	-.24**	1								
	N	253	253								
3. Dynamic training	r	-.72**	-.17**								
	N	253	253	253							
4. Dynamic transfer	r	.23**	.09	.21*	1						
	N	251	251	251	251						
Reading Achievement											
5. Reading Time 1	ρ	.06	.17*	-.19**	-.05	1					
	N	211	211	211	209	211					
6. Reading Time 2	ρ	.02	.20**	$-.14^{+}$	-.01	.66***	1				
	N	167	167	167	166	144	167				
7. Reading Time 3	ρ	.07	.01	$-.15^{+}$	-.13	.42**	.43**	1			
	N	141	141	141	141	121	131	123			
Math Achievement											
8. Math Time 1	ρ	.40**	.25**	-.53**	.24**	.47**	.24**	.27**	1		
	N	237	237	237	235	201	167	131	237		
9. Math Time 2	ρ	.33**	.23**	-.53**	.12	.44**	.33**	.23**	.78**	1	
	N	182	182	182	181	159	162	137	172	182	
10. Math Time 3	ρ	.32**	.07	-.45**	.06	.23*	.08	.37**	.60**	.64**	1
	N	138	138	138	138	118	120	120	137	138	138

$^{+}p < .10$, $^{*}p < .05$, $^{**}p < .01$, $^{***}p < .001$ (two-tailed)

Table 6.3 *Parameter estimates for fixed effects of probabilistic odds model with random intercepts for reading achievement prediction.*

Effect	Estimate	SE	*t*	Odds
Time	0.33	0.14	2.42*	1.39
Static pretest	0.10	0.14	0.70	1.11
Dynamic change	0.36	0.15	2.44*	1.43
Dynamic training	0.05	0.30	0.17	1.05
Dynamic transfer	-0.05	0.06	-0.79	0.95

$^{+}p < .10$, $^{*}p < .05$, $^{**}p < .01$, $^{***}p < .001$ (two-tailed)

Table 6.4 *Parameter estimates for fixed effects of probabilistic odds model with random intercepts for math achievement prediction.*

Effect	Estimate	SE	*t*	Odds
Time	-0.28	0.15	-1.86^{+}	0.76
Static pretest	0.02	0.23	0.10	1.02
Dynamic change	0.76	0.22	3.43***	2.14
Dynamic training	1.50	0.48	3.13**	4.48
Dynamic transfer	0.06	0.09	0.65	1.06

$^{+}p < .10$, $^{*}p < .05$, $^{**}p < .01$, $^{***}p < .001$ (two-tailed)

parameters. The variance of the random subjects effect subjects was: $\sigma^2 = .82$, $SE = .43$. This indicates moderate associations of reading achievement category for an individual across occasions. Of the fixed effects the dynamic change score and time were significant predictors of reading achievement. Neither the static pretest nor dynamic training and transfer scores explained additional variance in achievement category.

Math achievement

Table 6.4 presents the results of the model from Equation (1) for math achievement over time. This was also computed with the SAS NLMIXED procedure using Maximum Likelihood estimation comprising 554 observations (from 3 occasions for 205 students with 245 missing values). Model fit statistics were AIC = 822, BIC = 857, $-2LogLikelihood = 802$ with 10 parameters. The random effects have an estimated variance of $\sigma^2 = 3.37$, $SE = 1.07$, indicating strong within-subjects associations between occasions. The dynamic change score and dynamic training score were significant predictors of math achievement. Neither the static pretest nor dynamic transfer scores explained additional variance in achievement category.

6.4 Discussion

The main aim of this study was to investigate predictive value of dynamic testing outcomes on young children's school achievement in reading and math. We examined the predictive value of one static measure, the pretest score, and three dynamic test outcomes: performance change from pretest to posttest, training score and transfer score. The static measure, i.e. the figural analogies pretest, although related to math achievement, was surpassed as a correlate of achievement by the dynamic training measure, which refers to the amount of instruction an individual required to correctly solve the training items. The dynamic training measure was related to math achievement at each of the three time points, yet for reading achievement the association was only present in the same time period, but not for subsequent achievement. The dynamic transfer score was only related to concurrent math achievement. The dynamic measure of performance change and the training score (for math) provided the greatest predictive value of academic achievement over time.

Our findings are in line with previous research on the predictive validity of dynamic testing for math achievement, where performance change, the posttest and/or training scores are better or additional predictors of statically administered measures (Beckmann, 2006; Jeltova et al., 2011; Meijer, 1993; Resing, 1993; Tissink, Hamers, & Van Luit, 1993). For reading achievement, the associations between both static and dynamic figural analogy performance were not as strong as in previous studies (see review Caffrey et al., 2008). A dynamic test using a verbal task such as verbal analogies (Resing, 1993) or word decoding (D. F. Fuchs et al., 2011) may have produced stronger effects. However, performance change was a significant predictor of present and future reading achievement. The predictive power of our dynamic measures of analogical reasoning – especially performance change – above that of static measures confirmed our hypothesis and adds to the growing evidence of the predictive value of dynamic testing in psycho-educational assessment (e.g., Beckmann, 2006; Caffrey et al., 2008). Furthermore, we have demonstrated that the predictive value of dynamic testing may hold in longitudinal studies of typically developing children's performance on norm-referenced national achievement tests.

A disadvantage of choosing these national tests as achievement measures is that these are optional for schools which may lead to selection bias as the children were generally not measured at each time point on both reading and math. Furthermore, a different test, assessing slightly more advanced skills was used for each subsequent time point – therefore only progression relative to peers could be assessed and not growth in one particular subject area. In the future the predictive value of our dynamic measures should be assessed on latent scales of reading and math achievement that we can administer or obtain for more of our participants longitudinally.

Fluid reasoning ability is a well-established construct in cognitive ability testing (Freund & Holling, 2011a) and has been demonstrated to predict math and reading

achievement (e.g., Balboni et al., 2010; Ferrer & McArdle, 2004). Our finding that the dynamic measure of performance change is only somewhat related to initial fluid reasoning ability and appears to be an additional predictor of math and reading achievement, indicates that this may be a separate construct important in the assessment of learning and cognitive potential. Further research should focus on the predictive validity of the dynamic measure of performance change in other age groups and achievement domains to determine whether it indeed provides educators with a better picture of a child's capabilities and potential.

CHAPTER 7

General Discussion: Puzzling with potential – the bigger picture

The goal of this thesis project was to develop a new dynamic test of analogical reasoning for school children. The main aims of this thesis were to (1) investigate factors that influence children's differences in performance and change on this new dynamic test of analogical reasoning and (2) examine the predictive value of these dynamic measures on the children's school performance. In this final chapter first an introduction has been provided about AnimaLogica, the dynamic test of analogical reasoning we developed and report on throughout this thesis. In the following two sections investigations from previous chapters into the test design factors and person variables that may affect children's performance and change during dynamic testing have been discussed in reference to the literature. Finally, in section 4, we formulated general conclusions and address theoretical and practical implications.

7.1 AnimaLogica: A dynamic test of analogical reasoning for children

Dynamic testing was introduced in Chapter 1 as a means to measure children's potential for learning in developing cognitive abilities (Sternberg & Grigorenko, 2002). Measuring potential for learning is done by testing and training a child over one or multiple occasions. In AnimaLogica, as with its predecessor the Learning potential of Inductive Reasoning test (LIR, Resing, 1990, the training is provided in the form of graduated prompting techniques (Campione & Brown, 1987; Resing & Elliott, 2011). These interventions are incorporated into the training sessions that are preceded by a pretest and followed by a posttest: i.e., a pretest-training-posttest design. The *pretest* provides an indication of a child's *initial ability* in solving figural analogies (see Figure 7.1) and does not include training or feedback (Resing, 1997). The pretest is a form of static testing and is how conventional tests of cognitive abilities, such as an intelligence test, are usually administered. The pretest is followed by two *training* sessions in which the child receives the graduated

prompts training. Graduated prompting involves a standardized protocol of increasingly elaborate instructions starting with metacognitive prompts such as focusing attention, followed by cognitive prompts that explain the solving steps and ending with modeling with scaffolds where the trainer works through the problem step-by-step with the child (e.g., Bosma & Resing, 2012; Resing et al., 2009). An important aspect of the graduated prompts procedure is that instruction is only provided when the child is unable to solve the problem independently, thereby providing information on *instructional-needs*. The *number of prompts* required provides an indication of how much instruction a child needs to reach a particular performance level (Campione & Brown, 1987; Resing, 1993). The *type of prompts* that best aided solution – i.e. metacognitive, cognitive or modeling – may provide information on what type of instruction a child may benefit most from in future interventions (Resing, 2000). The training sessions are followed by a *posttest*, which is tailored instruction – i.e. *potential ability*. The *performance change* in the child's analogy solving from pretest to posttest shows how much can be learned from a short intervention. Examining the child's *self-explanations* and *solution strategies* provides information on the learning process – i.e. how an individual progressed during the dynamic test (e.g., Resing et al., 2009). The ability to solve and explain new but similar *transfer* problems may indicate the depth of learning an individual is capable of after a short, intensive training (e.g., Campione et al., 1985; Ferrara et al., 1986; Resing, 1990).

7.1.1 Main differences with earlier dynamic tests

In AnimaLogica, two problems that have prevented more wide-spread use of dynamic tests were addressed: (1) the extensive duration of administration and (2) the way learning and change is measured (Grigorenko & Sternberg, 1998). The administration of the test developed in this dissertation is considerably shorter than

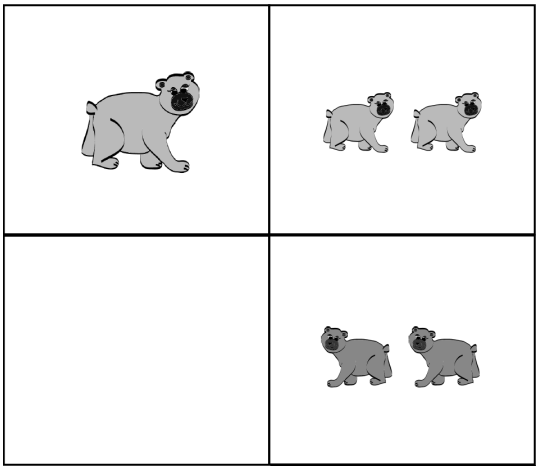

Figure 7.1 *An example figural analogy matrix item.*

previous ones – lasting approximately 80 minutes – and similar to traditional, static cognitive assessment batteries. This efficiency was achieved by providing a shorter training session and limiting assessment of performance on only one task, figural analogies, which could be easily implemented and administered on the computer (see Stevenson et al., 2011 for a discussion of paper versus computer administration). Secondly, the psychometric quality of dynamic tests is often unclear or considered poor as measuring *performance change* is often unreliable from the classical test theory perspective usually used in the statistical analyzes of dynamic tests. The main goal in the (ongoing) development of AnimaLogica was to keep it short and simple, while adhering to rigorous psychometric standards, yet still providing the valuable information unique to dynamic testing about an individual's learning process and cognitive potential.

7.1.2 Measurement Considerations

In the dynamic assessment literature, classical test theory measures tend to dominate (e.g., Calero et al., 2011; Resing et al., 2011; Tzuriel & Egozi, 2010). In the typical dynamic testing pretest-training-posttest design, often the posttest percentage correct scores are used as an indication of children's potential ability. However, gain scores (posttest minus pretest score) may be unreliable (Resing, Elliott, & Grigorenko, 2012). Another reason is that change is not necessarily measured on the same scale for test takers with different pretest scores – i.e. it is unlikely that an improvement of 4 correct items is the same if one had 3 or 16 items correct on the pretest. These problems with gains scores could potentially be solved when we use statistical models from item response theory (IRT). In the Rasch model, the simplest IRT model, the chance that an item is solved correctly depends on the difference between the test taker's latent ability 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), making IRT an appropriate method for measuring change (Embretson & Reise, 2000).

IRT measurement models for dynamic tests have gained some ground (e.g., Hessels & Bosson, 2003; De Beer, 2005). Embretson (1991b) extended the Rasch model and created the Multidimensional Rasch Model for Learning and Change (MRMLC). With this model it is possible to measure initial ability and modifiability (i.e. performance change) from one testing occasion to the next in a dynamic test, without the statistical pitfalls of classical test theory (Embretson, 1987, 1992). In the research with AnimaLogica reported in this thesis IRT models, the MRMLC and the mathematically similar Rasch model for repeated measurements developed by Andersen (Andersen, 1985), were used to measure pretest ability and performance change after training or posttest ability. In Chapter 5 we extended the MRMLC with

an explanatory component and thereby demonstrated the usefulness of De Boeck & Wilson's (De Boeck & Wilson, 2004) explanatory IRT approach in a dynamic testing context. Item response theory models hold 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.

AnimaLogica, as presented in this thesis, uses a non-adaptive item set for the pretest, training and posttest, where the pretest and posttest are isomorphs – i.e. the same problems but with different animals and colors. However, if older children or adults are to be tested then a larger difficulty range is required. In this case computer adaptive testing may be helpful, where the items of appropriate difficulty are selected or constructed from a large pool of possible items during testing (e.g., De Beer, 2005; Embretson, 2004). A downside of computer adaptive testing is that this would require more extensive data collection on item functioning prior to test development than was needed for the fixed item test we created.

A factor that certainly needs to be addressed in future research with AnimaLogica, and perhaps dynamic tests in general, is the scaling of the training items. Item response theory models such as the graded-response model (Samenjima, 1997) or partial credit model (Masters, 1982) seem appropriate for taking the number of required prompts or feedback interventions into consideration when estimating an individual's need for instruction during the training phase of the test (e.g., Attali, 2011; Wang & Heffernan, 2011). Furthermore, the dynamic Rasch model, which assesses whether learning has occurred during testing and the magnitude of individual differences in growth, may be also be appropriate (Verguts & De Boeck, 2000).

7.2 Factors affecting children's performance and change

Children's ability to solve figural analogies develops with great variability throughout childhood evidenced by large differences within each age group both in initial ability as well as performance change (e.g., Siegler & Svetina, 2002; Tunteler et al., 2008). There also appear to be considerable differences between children in the effects of retesting and training of figural analogies (Cheshire et al., 2005; Freund & Holling, 2011b). Similarly, dynamic testing studies show that children generally improve in analogy solving with training, interestingly again with large individual variation in improvement (e.g., Fabio, 2005; Jeltova et al., 2011). Dynamic tests aim to measure individual performance and change in order to gain insight into potential for learning. However, these differences in children's learning during dynamic testing appear to be influenced by test design factors such as training-type or item-format on the one hand and person variables such as working memory or ethnic background on the other hand. AnimaLogica has a number of possible diagnostic outcomes that could be influenced by test design factors: *initial ability, potential ability, performance change, instructional-needs, self-explanations, strategies* and *transfer.* Therefore, the research in this thesis investigated possible factors that could influence the measurement of children's potential for learning with AnimaLogica.

7.2.1 Test design factors

Although numerous aspects of the items or training format may influence children's performance on a dynamic test of analogical reasoning, this thesis was limited to address three of these factors: (1) training-type, (2) test item-format and (3) transfer task choice and administration. How each of these three factors affected children's AnimaLogica performance are now discussed in greater detail.

Much of the focus in the dynamic assessment literature is on "when" the training takes place (while testing or between test sessions) and "how" the training

is administered (standardized or not, individually or in a group) (Elliott, 2003; Sternberg & Grigorenko, 2002). AnimaLogica is an individually administered test using a standardized training within a pretest-training-posttest format. This type of dynamic testing format is often validated by comparing a group of trained children with a group that practices independently (e.g., Resing & Roth-Van Der Werf, 2003; Fabio, 2005) or a control condition in which the children receive regular classroom instruction (e.g., L. S. Fuchs et al., 2008) – thus serving as a control for retesting effects or general development. The research reported in this thesis demonstrated that AnimaLogica's graduated prompting training format was generally more effective in improving children's analogy solving than outcome-feedback training (Chapter 5), independent practice (Chapters 3 & 4) or control conditions (Chapters 2 & 3). The graduated prompts training of figural analogies was demonstrated to be an effective means of improving analogy solving with significant large effects comparable to that of other dynamic tests, despite the shorter duration (e.g., Resing, 2000). Furthermore, graduated prompting techniques, which include outcome and strategy-feedback as well as self-explanation prompts, appeared to provide children with more varied learning opportunities which resulted in greater potential results than outcome-feedback, practice or no training. This effect corresponds with work outside of dynamic testing such as the findings of Luwel et al. (Luwel et al., 2010) where strategy-feedback led to greater improvement in children's numerosity judgment than outcome-feedback. In the future, further validation of the strategy-based feedback component within the graduated prompts method could be assessed by comparing it with an outcome-feedback plus self-explanation condition.

A second test design factor investigated in this thesis was the role of item format. This factor has not received much attention in a dynamic testing literature, but as we demonstrated in Chapter 2, may be relevant in gaining insight into a

child's potential for learning. Multiple-choice items are often used in cognitive ability assessment, yet this may not be appropriate for dynamic testing as we were more interested in the problem solving process and not just if the child can select the correct answer option. We examined whether training during dynamic testing with multiple-choice or constructed-response items led to differences in children's analogy solving with regard to *strategy progression*, *self-explanation* or *performance change* from pretest to posttest. One group of children was trained on multiple-choice items (MC) and the second group was trained using constructed-response (CR) items – here they had to "construct" the answer in the empty box using a set of animal figures. The results did not show differences in *performance change* from pretest to posttest. The *number of prompts* the CR-trained children required was greater than that of the children in the MC-group, indicating that the CR-items were generally more difficult. Yet, children trained with CR-items provided better quality *self-explanations* compared to those trained with MC items. Also, a difference in *strategy progression* during training between the two training groups was apparent. Duplication is a commonly used strategy by young children who do not yet understand analogical reasoning; it refers to the answer being a copy of the figure in the adjoining box. This non-analogical strategy was used more often by the MC-group whereas the CR-group used a more advanced analogical strategy, partial correct. CR-items appeared to positively affect the children's understanding of analogical reasoning evidenced by better self-explanations and more advanced strategies, despite the greater difficulty of the items. This result coincided with other research in which more active processing has a greater learning effect (Harpaz-Itay et al., 2006; Martinez, 1999). Furthermore, CR-items provided more fine-grained analysis of the children's strategy-use and would therefore simplify diagnosis of erroneous reasoning (e.g., Birenbaum & Tatsuoka, 1987; Birenbaum et al., 1992). CR-items may be very beneficial for process-oriented diagnostics, with the goal of

adapting instruction to individual needs where the analysis of strategy progression and extent of understanding are of particular interest (e.g., Grigorenko, 2009a; Jeltova et al., 2007).

The third test design factor addressed in this thesis (Chapter 4) was which task can best be used to measure transfer and when to time the administration. Here we found that performance on the figural analogies pretest was strongly related to performance on the three possible transfer tasks we investigated: geometric analogies, seriation and analogy construction (e.g., Carpenter et al., 1990; Roth-Van Der Werf et al., 2002; Sternberg & Gardner, 1983). Yet, as with previous research on graduated prompting and transfer (e.g., Tunteler & Resing, 2010) we did not find differences in transfer between children trained with graduated prompts or those who practiced independently. Furthermore, children in both groups showed little improvement on the geometric analogies and seriation transfer tasks that were administered during the pretest and posttest sessions. Transfer is notoriously difficult to elicit in experimental settings (Barnett & Ceci, 2002). A possible explanation for our results and those of previous studies where training on a different task does not affect transfer of knowledge to similar tasks stems from Opfer and Thompson's (Opfer & Thompson, 2008) practice interference hypothesis. Their theory suggests that practice using incorrect solution strategies, which often occurs during pretesting, impedes transfer. This hypothesis was supported by the fact that transfer of analogical reasoning skills was only found to the reversal task, in which the child constructed an analogy for the examiner, which was not pretested. Reversal performance was related to *initial ability* on the figural analogies tasks, where more complex analogies were constructed by the children with higher pretest scores. The findings on the reversal task were in line with Siegler's theory (2006) that greater mastery of task strategies increases the chances of knowledge transfer to a novel situation in children. In the assessment of transfer within dynamic tests,

which often comprise a pretest-training-posttest format, it is perhaps advisable not to pretest the transfer tasks. Instead a selection of transfer tasks that measure similar skills to the tested task may provide more reliable measures. The effect of initial ability could be accounted for using the pretest scores of the trained task, which indeed correlated with performance on the analogy construction (reversal) task in the present study.

7.2.2 Person variables

Different dynamic tests have been developed with different populations in mind, from typically developing children (e.g., Resing & Elliott, 2011), intellectually or developmentally disabled persons (e.g., Hessels, 2009; Hessels-Schlatter, 2002) to clinical populations (e.g., Wiedl, Schöttke, Green, & Nuechterlein, 2004). AnimaLogica focuses on both typically developing elementary school children as reported in this thesis or those in a clinical educational setting (e.g., Resing, Bosma, & Stevenson, 2012). In both of these populations three so-called person variables are often reported in the literature that appear to influence children's performance and change on figural analogies: (1) cultural background, (2) working memory and (3) initial ability. The roles of each of these three factors in children's AnimaLogica performance were investigated in this thesis and are now discussed in greater detail.

Cultural background appears to play a role in performance on cognitive ability measures (e.g., Sternberg et al., 2007). For example, persons from the dominant culture generally obtain higher scores on measures of intelligence or reasoning ability (e.g., Te Nijenhuis & Van Der Vlier, 2001; Van de Vijver, 2002, 2008). These differences in conventional measures can be due to cultural bias in the tests themselves (i.e. item bias), the testing situation (e.g., nonnative instruction language, cultural influences on test-wiseness) or cultural differences in the tested construct (Grigorenko, 2009b; Van de Vijver & Poortinga, 1997). Dynamic testing

appears particularly valuable in groups that may be at a cultural disadvantage with traditional testing situations, such as ethnic minority populations, as the training opportunities can perhaps compensate for differences in test-wiseness or non-native instruction language (e.g., Hessels, 2000; Lidz & Pena, 1996; Tzuriel & Kaufman, 1999). Given our aim of developing a dynamic test that could easily be used in diagnostic practice it seemed imperative to consider the culturally diverse backgrounds of many school children in the Netherlands. Figural analogies were chosen as these are considered relatively culture-fair (Cattell, 1979). However, even such items may still be culturally biased (e.g., Van de Vijver, 2002). In Chapter 3 we examined the applicability of AnimaLogica in the dynamic testing of culturally diverse school populations in the Netherlands. In this study, the performance of 7-8 year old children with Dutch parents were compared to that of children with one or both parents from a different country (i.e. ethnic minority children). After confirming that the AnimaLogica items were not biased for one of the two groups, we investigated whether there were differences in their analogy solving progression during dynamic testing. Ethnicity was found to be related to *initial performance* on AnimaLogica as indigenous Dutch children obtained on average higher ability estimates on the pretest than ethnic minorities (e.g., Hamers et al., 1996; Tzuriel & Kaufman, 1999; Van de Vijver, 2002, 2008). However, no differences in *performance change* were found between indigenous and ethnic minority children. This result coincides with previous investigations into cultural differences on dynamic tests (Hamers et al., 1996; Tzuriel & Kaufman, 1999; Sternberg et al., 2007; Resing et al., 2009). Furthermore, we found that *instructional-needs* did not differ as both the number and type of required prompts during training were similar between the two groups. Also, the *self-explanations* of the indigenous Dutch and ethnic minority children did not differ. Cultural bias may still be present when ability is interpreted in the traditional sense as ethnic minorities have systematically lower

pre-test scores (Van de Vijver, 2008). However, dynamic measures, quantified by *performance change, self-explanations* and *instructional-needs*, did not appear to suffer from this bias. Dynamic testing may therefore potentially play a more prominent role in the culture-fair assessment of multicultural groups (Grigorenko, 2009b). Future investigations of AnimaLogica as an instrument for multicultural assessment should examine topics of cultural bias and equivalence in more depth.

A second factor that was investigated was working memory, which was addressed from different perspectives in Chapters 3, 4 and 5. Working memory refers to the ability to hold and manipulate entities in memory and shows large increases in childhood (e.g., Swanson, 2008). The role of age in analogy solving has been addressed in the earlier literature. Older children generally perform better on tests of analogical reasoning than younger children (e.g., Siegler & Svetina, 2002; Sternberg & Rifkin, 1979). In Chapter 5 we demonstrated that age is related to *initial ability* on the figural analogy problems, however this relation was confounded by working memory capacity. 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). We found that working memory capacity (WMC) was related to *initial ability* on AnimaLogica, whereby children with greater working memory had higher ability estimates. Yet, children with greater working memory efficiency did not profit more from graduated prompting than those with smaller working memory capacity – in other words working memory was unrelated to *performance change* in each of these studies. These results corroborate with those of Resing, Xenidou-Dervou, Steijn and Elliott (2012) in which the children also received graduated prompting on a different inductive reasoning task. The graduated prompts procedure provides step-by-step cognitive prompts of how to solve the tasks by attending to each transformation separately. A possible explanation for our findings is that this sequential approach

teaches the children a strategy to reduce the cognitive load of the task and thereby improve performance beyond that of control groups regardless of working memory efficiency. This idea is supported by our finding that lower WMC children required more cognitive prompts during the training yet improved their analogical reasoning to a similar extent as the children with higher WMC. A second possibility is that the graduated prompting procedure offers problem solving strategies or feedback that aids the children in more efficient use of their available working memory capacity. This is possibility seems supported by the results of Mackey, Stone, Hill and Bunge (2010) who found that performance on working memory tasks increased with an eight-week figural analogy training. However, in this case it concerns more intense training, therefore in future research WMC measures should be included both before and after training and help determine whether WM efficiency is affected by the graduated prompts intervention.

The third person variable that appears to play a role children's performance on a dynamic test is their initial ability – i.e. what they already know about solving analogies prior to training. We found that children with lower pretest scores generally improved more after the graduated prompts training than children with high *initial ability*, which given the moderate difficulty of the test items and the use of IRT estimations could not be due to ceiling effects (see Chapters 3 & 5). Our finding is in line with those of Swanson and Lussier's meta-analysis of dynamic testing effects who concluded that children with initially lower cognitive ability scores tend to improve more during short dynamic testing training-phases (Swanson & Lussier, 2001). Furthermore, in training studies outside of the dynamic testing domain similar results are found. In the case of Luwel et al. (2010) children with lower intelligence test scores improved more with strategy-feedback compared to children with high intelligence scores. Also, Jaeggi et al. (2008) found that low ability children tended to improve more so than high ability children on figural

matrices after training on a working memory task. This finding 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 reasoning test. It also appears that the IRT-based measure of *performance change* is more suitable in identifying these children than measures of *instructional-needs* as the number of required prompts in training correlates more strongly with initial ability than with performance change (see Chapter 6).

7.3 Predictive value

The final puzzle piece we investigated was whether recent school performance was related to analogy solving and improvement during dynamic testing. The main aim of Chapter 6 was to investigate predictive value of dynamic testing outcomes on young children's school achievement in reading and math. Dynamic measures may provide additional predictive value of school achievement in reading (e.g., Bynre et al., 2000; D. F. Fuchs et al., 2011; Swanson, 2011b) and math (e.g., Beckmann, 2006; Jeltova et al., 2011; Resing, 1993; Sittner Bridges & Catts, 2011). However, dynamic testing studies do not consistently show advantages of dynamic measures in predicting achievement (e.g., Caffrey et al., 2008). Furthermore, a variety of dynamic measures have been used to predict achievement and it is unclear which dynamic measure (e.g., *potential ability, performance change, instructional-needs, transfer*) is most useful. We compared the predictive value of AnimaLogica'static measure, the pretest score, to three dynamic measures: performance change, instructional-needs and transfer score. The static measure, i.e. the figural analogies pretest, was strongly associated with math achievement, but was surpassed as a correlate of achievement by *instructional-needs* – i.e. the amount of instruction the child needed to correctly solve the training items. In Chapter 3 we had already seen that the children's instructional-needs correlated strongly with teacher ratings and learning

ability – which may mean we are tapping into similar information the teacher obtains in the classroom on individual children's ability to learn from instruction (Bosma & Resing, 2012). Yet, instructional-needs and the children's transfer score from the reversal task were often more strongly related to academic achievement measured in the same time period, but not necessarily to subsequent achievement. The dynamic measure of *performance change* provided additional predictive value of reading and math achievement over the course of three measures within one year. This result coincides with Freund and Holling's (2011b) finding that children with higher school grades show the greatest improvement upon retesting. Furthermore, our findings were in line with previous research on the predictive validity of dynamic testing, where performance change, the posttest and/or training scores are better or additional predictors of statically administered measures (Beckmann, 2006; Jeltova et al., 2011; Resing, 1993). The unique contribution of this study was the longitudinal design in which future rather than concurrent achievement was predicted and the identification of which of the dynamic measures provide the best prediction.

7.4 Conclusion

On the whole, children showed great variation in their potential for learning to solve analogies. As with previous research on children's analogy solving progression, the children's performance generally improved over repeated testing occasions, but the degree of improvement varied greatly (e.g., Freund & Holling, 2011b; Siegler & Svetina, 2002; Tunteler & Resing, 2007c, 2007b). The large individual differences in performance and change after the short dynamic testing 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). In AnimaLogica this variation was present in

each of the investigated dynamic measures: *strategy-progression*, *self-explanations*, *performance change*, *instructional-needs* and *transfer.* The type of training influenced each of these measures of AnimaLogica performance (Chapters 2 - 5). Also, the item format affected *performance change*, *strategy-progression*, *self-explanations* and *instructional-needs* (Chapter 2). *Transfer* performance was related to initial ability and working memory (Chapter 4). Yet, the person variables we investigated, ethnicity and working memory, were not related to *performance change* (Chapters 3, 4 & 5). With regard to ethnicity this technically negative finding is in fact positive as similar dynamic outcomes (*performance change*, *self-explanations* and *instructional-needs*) between indigenous Dutch and ethnic minority children seems to indicate that AnimaLogica may be an appropriate measure for culturally diverse school children (Chapter 3).

However, given the importance placed upon working memory in cognitive and psychoeducational assessment (e.g., Pickering & Gathercole, 2004) it was important to investigate whether working memory could explain children's differences in the *performance change* and *transfer* on our dynamic test. We found working memory was unrelated to both aspects. Performance change and ability to transfer knowledge to novel situations, such as in the reversal task, are not often included in the assessment of intellectual abilities (Bosma & Resing, 2006; Elliott et al., 2010), yet the findings in this thesis indicate that these two dynamic measures may be separate constructs and important in the assessment of learning and cognitive potential.

Initial ability does seem to affect how children progress in analogy solving during dynamic testing. For example, higher ability children generally require fewer prompts (Chapters 3 & 6) and show greater *transfer* on the reversal task (Chapter 4). Yet, lower ability children tended to show greater *performance change* (Chapter 5). This finding is important because it demonstrates that the children with untapped potential are most likely to be found at the lower end of the spectrum

of static testing scores (e.g., Swanson & Lussier, 2001).

The predictive power of our dynamic measures of analogical reasoning – especially Rasch-scaled performance change – above that of static measures confirmed our hypothesis and adds to the growing evidence of the predictive value of dynamic testing in psycho-educational assessment (Chapter 6). Analogical reasoning is often measured in cognitive ability tests (Freund & Holling, 2011a) and has been demonstrated to predict math and reading achievement (e.g., Balboni et al., 2010). Our finding that the dynamic measure of performance change is only somewhat related to *initial ability* and appears to be a better predictor of math and reading achievement, provides further evidence that this may be a separate construct important in the assessment of learning and cognitive potential. Furthermore, the performance change measure, which has often been criticized as a measure of learning potential in the context of classical test theory (e.g., Sternberg & Grigorenko, 2002), has demonstrated its worth when estimated using item response theory models and will hopefully find its place again among the valuable measurement outcomes of potential for learning.

AnimaLogica outcomes appear to be a valuable addition to conventional tests in the prediction of scholastic achievement and applicable for culturally diverse school populations. Furthermore, process-oriented diagnostic information, such as *performance change, instructional-needs, self-explanations, strategies* and *transfer* are available. This information may prove useful for educators in providing interventions that help children more thoroughly utilize their potential for learning at school (e.g., Bosma & Resing, 2012; Jeltova et al., 2011).

References

Ackerman, P., Beier, M., & Boyle, M. (2005). Working memory and intelligence: The same or different constructs? *Psychological Bulletin*, *131*(1), 30-60.

Alexander, P. A., Willson, V. L., White, C. S., & Fuqua, J. D. (1987). Analogical reasoning in young children. *Journal of Educational Psychology*, *79*(4), 401-408.

Alloway, T. P. (2007). *Automated working memory assessment (awma)*. London: Harcourt Assessment.

Alloway, T. P., Gathercole, S. E., & Pickering, S. J. (2006). Verbal and visuospatial short-term and working memory in children: Are they separable? *Child Development*, *77*(6), 1698-1716.

Alloway, T. P., Gathercole, S. E., Willis, C., & Adams, A. (2004). A structural analysis of working memory and related cognitive skills in young children. *Journal of Experimental Child Psychology*, *87*(2), 85-106.

Andersen, E. B. (1985). Estimating latent correlations between repeated testings. *Psychometrika*, *50*(1), 3-16.

Attali, Y. (2011). Immediate feedback and opportunity to revise answers: Application of graded-response IRT model. *Applied Psychological Measurement*, *35*(6), 472-479.

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

Bacon, A. M., Handley, S. J., Dennis, I., & Newstead, S. E. (2008). Reasoning strategies: The role of working memory and verbal-spatial ability. *European Journal of Cognitive Psychology*, *20*(6), 1065-1086.

Baddeley, A. D., & Hitch, G. (1974). Working memory. In G. H. Bower (Ed.), *The psychology of learning and motivation* (p. 47-90). New York: Academic Press.

Balboni, G., Naglieri, J. A., & Cubelli, R. (2010). Concurrent and predictive validity of the progressive matrices and the naglieri nonverbal ability test. *Journal of Psychoeducational Assessment*, *28*(3), 222-235.

Ball, L. J., Hoyle, A. M., & Towse, A. S. (2010). The facilitatory effect of negative feedback on the emergence of analogical reasoning abilities. *British Journal of Developmental Psychology*, *28*(3), 583-602.

Barnett, S. M., & Ceci, S. J. (2002). When and where do we apply what we learn? A taxonomy for far transfer. *Psychological Bulletin*, *128*(4), 612-637.

Bates, D., & Maechler, M. (2010). *lme4: Linear mixed modeling using S4 classes.* (Computer program and manual). Available from `http://cran.r-project.org/web/packages/lme4/index.html`.

Beckmann, J. F. (2006). Superiority: Always and everywhere? On some misconceptions in the validation of dynamic testing. *Educational and Child Psychology*, *23*(3), 35-49.

Behuniak, P., Rogers, J. B., & Dirir, M. A. (1996). Item function characteristics and dimensionality for alternative response formats in mathematics. *Applied Measurement in Education*, *9*(3), 257-275.

Bereiter, C. (1963). Some persisting dilemmas in the measurement of change. In I. C. W. Harris (Ed.), *Problems in measuring change* (p. 3-20). Madison: University of Wisconsin Press.

Bernardo, A. B. I. (2001). Analogical problem construction and transfer in mathematical problem solving. *Educational Psychology*, *21*(2), 137-150.

Bethel-Fox, C. E., Lohman, D. F., & Snow, R. E. (1984). Adaptive reasoning: Componential and eye movement analysis of geometric analogy performance. *Intelligence*, *8*(3), 205-238.

Beunher, M., Krumm, S., & Pick, M. (2005). Reasoning = working memory ≠ attention. *Intelligence*, *33*(3), 251-272.

Birenbaum, M., & Tatsuoka, K. K. (1987). Open-ended versus multiple-choice item formats - it does make a difference for diagnostic purposes. *Applied Psychological Measurement*, *11*(4), 385-395.

Birenbaum, M., Tatsuoka, K. K., & Gutvirtz, Y. (1992). Effects of response format on diagnostic assessment of scholastic achievement. *Applied Psychological Measurement*, *16*(4), 353-363.

Bleichrodt, N., Drenth, P. J. D., Zaal, J. N., & Resing, W. C. M. (1987). *Handleiding bij de revisie amsterdamse kinder intelligentie test [manual of the revised amsterdam child intelligence test]*. Lisse: Swets and Zeitlinger.

Bosma, T., & Resing, W. C. M. (2006). Dynamic assessment and a reversal task: A contribution to needs-based assessment. *Educational and Child Psychology*, *23*(3), 81-98.

Bosma, T., & Resing, W. C. M. (2012). Need for instruction: Dynamic testing in special education. *European Journal of Special Needs Education*, *27*(1), 1-19.

Bosma, T., Stevenson, C. E., & Resing, W. C. M. (submitted). Differences in need for instruction: dynamic testing in children with arithmetic difficulties.

Bransford, J. D., & Schwartz, D. L. (1999). Rethinking transfer: A simple proposal with multiple implications. *Review of Research in Education*, *24*, 61-100.

Bridgeman, B. (1992). A comparison of quantitative questions in open-ended and multiple-choice formats. *Journal of Educational Measurement*, *29*, 253-271.

Bridgeman, B., & Buttram, J. (1975). Race differences on nonverbal analogy test performance as a function of verbal strategy training. *Journal of Educational*

Psychology, *67*, 586-590.

Brown, A. L., & French, L. A. (1979). The zone of potential development: implications for intelligence testing in the year 2000. *Intelligence*, *3*(3), 255-271.

Brown, A. L., & Kane, M. (1988). Preschool children can learn to transfer: Learning to learn and learning from example. *Cognitive Psychology*, *20*(4), 493-523.

Bynre, B., Fielding-Barnsley, R., & Ashley, L. (2000). Effects of preschool phoneme identity training after six years: Outcome level distinguished from rate of response. *Journal of Mental Deficiency*, *76*, 159-169.

Caffrey, E., Fuchs, D., & Fuchs, L. S. (2008). The predictive validity of dynamic assessment: A review. *Journal of Special Education*, *41*(4), 254-270.

Calero, M. D., Belen, G. M., & Robles, M. A. (2011). Learning potential in high IQ children: The contribution of dynamic assessment to the identification of gifted children. *Learning and Individual Differences*, *21*(2), 176-181.

Campione, J. C., & Brown, A. L. (1987). Linking dynamic assessment with school achievement. In C. S. Lidz (Ed.), *Dynamic assessment: an interactional approach to evaluating learning potential* (p. 82-109). New York: Guilford Press.

Campione, J. C., Brown, A. L., Ferrara, R., Jones, R., & Steinberg, E. (1985). Breakdowns in flexible use of information: Intelligence-related differences in transfer following equivalent learning performance. *Intelligence*, *9*, 297-315.

Carlson, J. S., & Wiedl, K. H. (1992). Principles of dynamic assessment: The application of a specific model. *Learning and Individual Differences*, *4*(2), 153-166.

Carpenter, P. A., Just, M. A., & Shell, P. (1990). What one intelligence test measures: a theoretical account of the processing in the Raven Progressive Matrices test. *Psychological Review*, *97*(3), 404-431.

Cattell, R. B. (1979). Are culture fair intelligence tests possible and necessary? *Journal of Research and Development in Education*, *12*(2), 3-13.

Cheshire, A., Ball, L. J., & Lewis, C. (2005). Self-explanation, feedback and

the development of analogical reasoning skill: Microgenetic evidence for a metacognitive processing account. In B. G. Bara, L. Barsalou, & M. Bucciarelli (Eds.), *Perspectives on thought and language: Interrelations in development* (p. 435-440). New Jersey: Lawrence Erlbaum Associates, Inc.

Cho, S., Holyoak, K. J., & Cannon, T. D. (2007). Analogical reasoning in working memory: resources shared among relational integration, interference resolution, and maintenance. *Memory and Cognition*, *35*(6), 1445-1455.

CITO. (2010a). *Leerling- en onderwijsvolgsysteem, ordenen, groep 1/2* [Monitoring and evaluation system for primary pupils – Mathematical Reasoning, kindergarten]. Arnhem, The Netherlands: Author.

CITO. (2010b). *Leerling- en onderwijsvolgsysteem, rekenen-wiskunde, groep 3* [Monitoring and evaluation system for primary pupils – Arithmetic and Mathematics, grade 1]. Arnhem, The Netherlands: Author.

CITO. (2010c). *Leerling- en onderwijsvolgsysteem, rekenen-wiskunde, groep 4* [Monitoring and evaluation system for primary pupils – Arithmetic and Mathematics, grade 2]. Arnhem, The Netherlands: Author.

CITO. (2010d). *Leerling- en onderwijsvolgsysteem, drie minuten toets, groep 3* [Monitoring and evaluation system for primary pupils – Arithmetic and Reading, grade 1]. Arnhem, The Netherlands: Author.

Conway, A. R. A., Cowan, N., Bunting, M. F., Therriault, D. J., & Minkoff, S. R. B. (2002). A latent variable analysis of working memory capacity, short-term memory capacity, processing speed, and general fluid intelligence. *Intelligence*, *30*, 163-183.

Coventry, W. L., Byrne, B., Olsen, R. K., Corley, R., & Samuelsson, S. (2011). Dynamic and static assessment of phonological awareness in preschool: A behavior-genetic study. *Journal of Learning Disabilities*, *44*(4), 322-329.

Csapó, B. (1997). The development of inductive reasoning: Cross-sectional assessments in an educational context. *International Journal of behavioral*

development, *20*(4), 609-626.

Currie, M., & Chiramanee, T. (2010). The effect of multiple-choice item format on the measurement of knowledge of language structure. *Language Testing*, *27*(4), 471-491.

Dahlin, E., Neely, A. S., Larsson, A., Bäckmann, L., & Nyberg, L. (2008). Transfer of learning after updating training mediated by striatum. *Science*, *320*, 1510-1512.

Day, J. D., Engelhardt, J. L., Maxwell, S. E., & Bolig, E. E. (1997). Comparison of static and dynamic assessment procedures and their relation to independent performance. *Journal of Educational Psychology*, *89*, 358-368.

De Beer, M. (2005). Development of the learning potential computerized adaptive test (LPCAT). *South African Journal of Psychology*, *35*(4), 717-747.

De Bock, D. F. (1976). Basic issues in the measurement of change. In D. de Gruijter & L. Van der Kamp (Eds.), *Advances in psychological and educational measurement* (p. 75-96). New York: Wiley.

De Boeck, P. A. L. (2008). Random item IRT models. *Psychometrika*, *73*(4), 533-559.

De Boeck, P. A. L., Bakker, M., Zwitser, R., Nivard, M., Hofman, A., Tuerlinckx, F., et al. (2011). The estimation of item response models with the lme4. *Journal of Statistical Software*, *39*(12), 1-28.

De Boeck, P. A. L., & Wilson, M. (2004). *Explanatory item response models: A generalized linear and nonlinear approach*. New York: Springer.

De Boeck, P. A. L., Wilson, M., & Acton, S. (2005). A conceptual and psychometric framework for distinguishing categories and dimensions. *Psychological Review*, *112*(1), 129-158.

De Corte, E. (2003). Transfer as the productive use of acquired knowledge, skills and motivations. *Current Directions in Psychological Science*, *12*(4), 142-146.

De Smedt, B., Janssen, R., Bouwens, K., Verschaffel, L., Boets, B., & Ghesquière, P. (2009). Working memory and individual differences in math achievement: A

longitudinal study from first grade to second grade. *Journal of Experimental Child Psychology*, *103*(2), 186-201.

Detterman, R. J. (1993). The case for prosecution: Transfer as an epiphenomenon. In D. K. Detterman & R. J. Sternberg (Eds.), *Transfer on trial: Intelligence, cognition, and instruction* (p. 1-24). Norwood, NJ: Ablex Publishing.

Dörfler, T., Golke, S., & Artelt, C. (2009). Dynamic assessment and its potential for assessment of reading competence. *Studies in Educational Evaluation*, *35*, 77-82.

Durost, W. N., Gardner, E. F., & Madden, R. (1970). *Analysis of learning potential*. New York: Harcourt, Brace and World.

Elliott, J. G. (2003). Dynamic assessment in educational settings: Realising potential. *Educational Review*, *55*(1), 15-32.

Elliott, J. G., Grigorenko, E. L., & Resing, W. C. M. (2010). Dynamic assessment: The need for a dynamic approach. In P. Peterson, E. Baker, & B. M. (Eds.) (Eds.), *International encyclopedia of education* (Vol. 3, p. 220-225). Amsterdam: Elsevier.

Embretson, S. E. (1987). Improving the measurement of spatial aptitude by dynamic testing. *Intelligence*, *11*, 333-358.

Embretson, S. E. (1991a). Implications of a multidimensional latent trait model for measuring change. In L. M. Collins & J. L. Horn (Eds.), *Best methods for the analysis of change* (p. 184-203). Washington, DC: American Psychological Association.

Embretson, S. E. (1991b). A multidimensional latent trait model for measuring learning and change. *Psychometrika*, *56*(3), 495-515.

Embretson, S. E. (1992). Measuring and validating cognitive modifiability as an ability: A study in the spatial domain. *Journal of Educational Measurement*, *29*(1), 25-50.

Embretson, S. E. (2004). Measuring human intelligence with artificial intelligence: Adaptive item generation. In R. J. Sternberg & J. Pretz (Eds.), *Cognition and intelligence* (p. 251-267). New York: Cambridge University Press.

Embretson, S. E., & Prenovorst, L. K. (2000). Dynamic cognitive testing: What kind of information is gained by measuring response time and modifiability? *Educational and Psychological Measurement*, *60*(6), 837-863.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Erlbaum Publishers.

Engel de Abreu, P. M. J., Conway, A. R. A., & Gathercole, S. E. (2010). Working memory and fluid intelligence in young children. *Intelligence*, *38*, 552-561.

Fabio, R. A. (2005). Dynamic assessment of intelligence is a better reply to adaptive behavior and cognitive plasticity. *The Journal of General Psychology*, *132*, 41-64.

Facon, B., Magis, D., Nuchadee, M., & Boeck, P. A. L. de. (2011). Do raven's colored progressive matrices function in the same way in typical and clinical populations? Insights from the intellectual disabilities field. *Intelligence*, *39*, 281-291.

Fagan, J. F., & Holland, C. R. (2007). Racial equality in intelligence: Predictions from a theory of intelligence as processing. *Intelligence*, *35*, 319-334.

Fagan, J. F., & Holland, C. R. (2009). Culture-fair prediction of academic achievement. *Intelligence*, *37*, 62-67.

Ferrara, R. A., Brown, A. L., & Campione, J. C. (1986). Children's learning and transfer of inductive reasoning rules: Studies of proximal development. *Child Development*, *57*, 1087-1099.

Ferrer, E., & McArdle, J. J. (2004). An experimental analysis of dynamic hypotheses about cognitive abilities and achievement from childhood to early adulthood. *Developmental Psychology*, *40*(6), 935-952.

Ferrer, E., McArdle, J. J., Shaywitz, B. A., Holanhan, J. M., Marchione, K., & Shaywitz, S. E. (2007). Longitudinal models of developmental dynamics between reading and cognition from childhood to adolescence. *Developmental Psychology*, *43*(6), 1460-1473.

Freedle, R. (2003). Correcting the SAT's ethnic and social-class bias: A method for

reestimating SAT scores. *Harvard Educational Review*, *73*(1), 1-43.

Freund, P. A., & Holling, H. (2011a). How to get really smart: Modeling retest and training effects in ability testing using computer-generated figural matrix items. *Intelligence*, *39*, 233-243.

Freund, P. A., & Holling, H. (2011b). Retest effects in matrix test performance: Differential impact of predictors at different hierarchy levels in an educational setting. *Learning and individual differences*, *21*(5), 597-601.

Fry, A. F., & Hale, S. (1996). Processing speed, working memory and fluid intelligence: Evidence for a developmental cascade. *Psychological Science*, *7*(4), 237-241.

Fuchs, D. F., Compton, D. L., Fuchs, L. S., Bouten, B., & Caffrey, E. (2011). The construct and predictive validity of a dynamic assessment of young children learning to read: Implications for rti-frameworks. *Journal of Learning Disabilities*, *44*(4), 339-347.

Fuchs, L. S., Compton, D. L., Fuchs, D. F., Hollenbeck, K. N., Craddock, C. F., & Hamlett, C. L. (2008). Dynamic assessment of algebraic learning in predicting third graders' development of mathematical problem solving. *Journal of Educational Psychology*, *100*(4), 829-850.

Gathercole, S. E., Pickering, B., S. J. Ambridge, & Wearing, H. (2004). The structure of working memory from 4 to 15 years of age. *Journal of Developmental Psychology*, *40*, 177-190.

Gay, L. R. (1980). The comparative effects of multiple-choice versus short-answer tests on retention. *Journal of Educational Measurement*, *17*, 45-50.

Goswami, U. (1992). *Analogical reasoning in children*. Hove, UK: Lawrence Erlbaum Associates.

Grigorenko, E. L. (2009a). Dynamic assessment and response to intervention: two sides of one coin. *Journal of Learning Disabilities*, *42*, 111-132.

Grigorenko, E. L. (2009b). *Multicultural psychoeducational assessment*. New York: Springer Publishing Company LLC.

Grigorenko, E. L., & Sternberg, R. J. (1998). Dynamic testing. *Psychological Bulletin*, *124*(1), 75-111.

Hager, W., & Hasselhorn, M. (1998). The effectiveness of the cognitive training for children from a differential perspective. *Learning and Instruction*, *8*, 411-438.

Hamers, J. H. M., Hessels, M. G. P., & Pennings, A. H. (1996). Learning potential in ethnic minority children. *European Journal of Psychological Assessment*, *12*(3), 183-192.

Harpaz-Itay, Y., Kaniel, S., & Ben-Amram, E. (2006). Analogy construction versus analogy solution, and their influence on transfer. *Learning and Instruction*, *16*, 583-591.

Hatcher, J., Snowling, M. J., & Griffiths, Y. M. (2002). Cognitive assessment of dyslexic students in higher education. *British Journal of Educational Psychology*, *72*(1), 119-133.

Haywood, H. C., & Lidz, C. S. (2007). *Dynamic assessment in practice: Clinical and educational applications*. New York: Cambridge University Press.

Hedden, T., Park, D. C., Nisbett, R., Ji, L., Jing, Q., & Jiao, S. (2002). Cultural variation in verbal versus spatial neuropsychological function across the life span. *Neuropsychology*, *16*(1), 65-73.

Helms-Lorenz, M., & Van de Vijver, F. J. R. (1995). Cognitive assessment in education in a multicultural society. *European Journal of Psychological Assessment*, *11*(3), 158-169.

Helms-Lorenz, M., Van de Vijver, F. R., & Poortinga, Y. H. (2003). Cross-cultural differences in cognitive performance and spearman's hypothesis: g or c? *Intelligence*, *31*(1), 9-29.

Hessels, M. G. P. (2000). The learning potential test for ethnic minorities (LEM): A

tool for standardized assessment of children in kindergarten and the first years of primary school. In C. S. Lidz & J. G. Elliott (Eds.), *Dynamic assessment: Prevailing models and applications* (p. 109-131). Oxford: Elsevier Inc.

Hessels, M. G. P. (2009). Estimation of the predictive validity of the HART by means of a dynamic test of geography. *Journal of Cognitive Education and Psychology*, *8*(1), 5-21.

Hessels, M. G. P., & Bosson, M. O. (2003). *Hessels Analogical Reasoning Test (HART): Instruction manual (unpublished)*. Faculty of Psychology and Educational Sciences: University of Geneva.

Hessels-Schlatter, C. (2002). A dynamic test to assess learning capacity in people with severe impairments. *American Journal on Mental Retardation*, *107*(5), 340-351.

Hickendorff, M., Heiser, W. J., Van Putten, C., & Verhelst, N. D. (2008). Solution strategies and achievement in dutch complex arithmetic: Latent variable modeling of change. *Psychometrika*, *74*(2), 331-350.

Hickendorff, M., Van Putten, C. M., Verhelst, N. D., & Heiser, W. J. (2010). Individual differences in strategy use on division problems: Mental versus written computation. *Journal of Educational Psychology*, *102*(2), 438-452.

Hornung, C., Brunner, M., Rueter, R. A. P., & Martin, R. (2011). Children's working memory: Its structure and relationship to fluid intelligence. *Intelligence*, *39*, 210-221.

Hosenfeld, D., B Van den Boom, & Resing, W. C. M. (1997). Constructing geometric analogies for the longitudinal testing of elementary school children. *Journal of Educational Measurement*, *34*, 367-372.

In'nami, Y., & Kozumi, R. (2009). A meta-analysis of test format effects on reading and listening test performance: Focus on multiple-choice and open-ended formats. *Language Testing*, *26*(2), 219-244.

Jaarsveld, S., Lachmann, T., & Van Leeuwen, C. (2012). Creative reasoning

across developmental levels: Convergence and divergence in problem creation. *Intelligence*, *40*(2), 172-188.

Jacobs, P. J., & Vandeventer, M. (1971). The learning and transfer of double-classification skills: A replication and extension. *Journal of Experimental Child Psychology*, *42*, 240-257.

Jaeggi, S. M., Buschkuehl, M., J., J., & Perrig, W. J. (2008). Improving fluid intelligence with training on working memory. *Proceedings of National Academy of Science*, *105*(19), 6829-6833.

Janssen, R., De Boeck, P. A. L., Viane, M., & Vallaeys, L. (1999). Simple mental addition in children with and without mild mental retardation. *Journal of Experimental Child Psychology*, *74*, 261-281.

Janssen, R., Schepers, J., & Peres, D. (2004). Models with item group parameters. In I. P. A. L. De Boeck & M. Wilson (Eds.), *Explanatory item response models: a generalized linear and nonlinear approach* (p. 189-212). New York: Springer.

Jeltova, I., Birney, D., Fredine, N., Jarvin, L., Sternberg, R. J., & Grigorenko, E. L. (2007). Dynamic assessment as process-oriented assessment in educational settings. *International Journal of Speech-Language Pathology*, *9*(4), 273-285.

Jeltova, I., Birney, D., Fredine, N., Jarvin, L., Sternberg, R. J., & Grigorenko, E. L. (2011). Making instruction and assessment responsive to diverse students' progress: group-administered dynamic assessment in teaching mathematics. *Journal of Learning Disabilities*, *44*(4), 381-395.

Kail, R. V. (2007). Longitudinal evidence that increases in processing speed and working memory enhance children's reasoning. *Association for Psychological Science*, *18*, 312-313.

Klauer, K. J., & Phye, G. D. (2008). Inductive reasoning: A training approach. *Review of Educational Research*, *78*(1), 85-123.

Krumm, M., S Ziegler, & Buehner, M. (2008). Reasoning and working memory as

predictors of school grades. *Learning and Individual Differences*, *18*, 248-257.

Kyllonen, P. C., & Christal, R. E. (1990). Reasoning ability is (little more than) working-memory capacity?! *Intelligence*, *14*, 389-433.

Leech, R., Mareschal, D., & Cooper, R. P. (2008). Analogy as relational priming: A developmental and computational perspective on the origins of a complex cognitive skill. *Behavioral and Brain Sciences*, *31*, 357-414.

Lidz, C. S., & Pena, E. (1996). Dynamic assessment: The role, its relevance as a non-biased approach, and its application to Latino American children. *Language, Speech, and Hearing in the Schools*, *27*, 367-372.

Lievens, F., Reeve, C. L., & Heggestad, E. D. (2007). An examination of psychometric bias due to retesting on cognitive ability tests in selection settings. *Journal of Applied Psychology*, *92*, 1672-1682.

Lifshitz, H., Tzuriel, D., & Weiss, I. (2005). Effects of training in conceptual versus perceptual analogies among adolescents and adults with intellectual disability. *Journal of Cognitive Education and Psychology*, *5*(2), 144-167.

Lord, F. M. (1963). Elementary models for measuring change. In C. W. Harris (Ed.), *Problems in measuring change* (p. 21-38). Madison: University of Wisconsin Press.

Luo, D., Thompson, L., & Detterman, R. J. (2003). The causal factor underlying the correlation between psychometric g and scholastic performance. *Intelligence*, *31*(1), 67-83.

Luwel, K., Foustana, A., Papadatos, Y., & Verschaffel, L. (2010). The role of intelligence and feedback in children's strategy competence. *Journal of Experimental Child Psychology*, *108*, 61-76.

Mackey, A. P., Hill, S. S., Stone, S. I., & Bunge, S. A. (2010). Differential effects of reasoning and speed training in children. *Developmental Science*, *14*(3), 582-290.

Magis, D., Béland, S., Tuerlinckx, F., & Boeck, P. A. L. de. (2010). A general framework and an r package for the detection of dichotomous differential item

functioning. *Behavior Research Methods*, *42*(3), 847-862.

Martinez, M. E. (1999). Cognition and the question of item format. *Educational Psychologist*, *34*(4), 207-218.

Martinussen, R., Hayden, J., Hogg-Johnson, S., & Tannock, R. (2005). A meta-analysis of working memory impairments in children with attention-deficit/hyperactivity disorder. *Journal of American Academy of Child and Adolescent Psychiatry*, *44*(4), 377-384.

Masters, G. N. (1982). A rasch model for partial credit scoring. *Psychometrika*, *47*(2), 149-172.

Meijer, J. (1993). Learning potential, personality characteristics, and test performance. In J. H. M. Hamers, K. Sijtsma, & A. J. J. M. Ruijssenaars (Eds.), *Learning potential assessment: Theoretical, methodological and practical issues* (p. 341-362). Lisse: Swets and Zeitlinger, Inc.

Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika*, *4*, 525-543.

Miller, P. H., & Seier, W. L. (1994). Strategy utilization deficiencies in children: When, where, and why. In H. W. Reese (Ed.), *Advances in child development and behavior* (p. 107-156). 25, . New York: Academic Press.

Millsap, R. (2010). Testing measurement invariance using item response theory in longitudinal data: An introduction. *Child Development Perspectives*, *4*(5-9).

Molenberghs, G., & Verbeke, G. (2005). *Models for discrete longitudinal data*. New York: Springer.

Morrison, R. G., Holyoak, K. J., & Truong, B. (2001). Working-memory modularity in analogical reasoning. In *Proceedings of the twentyfourth annual conference of the cognitive science society* (p. 663-668). Mahwah, NJ: Lawrence Erlbaum.

Mulholland, T. M., Pellegrino, J. W., & Glaser, R. (1980). Components of geometric analogy solution. *Cognitive Psychology*, *12*, 252-284.

Neisser, U., Boodoo, G., Bouchard, T. J. j., Boykin, A. W., Brody, N., Halpern, D. F., et al. (1996). Intelligence: Knowns and unknowns. *American Psychologist*, *51*(2), 77-101.

Opfer, J. E., & Thompson, C. A. (2008). The trouble with transfer: Insights from microgenetic changes in the representation of numerical magnitude. *Child Development*, *79*(3), 788-804.

Pena, E., Iglasius, A., & Lidz, C. S. (2001). Reducing test bias through dynamic assessment of children's word learning abilitiy. *American Journal of Speech-Language Pathology*, *10*, 138-154.

Pickering, S. J., & Gathercole, S. E. (2004). Distinctive working memory profiles in children with special educational needs. *Educational Psychology*, *24*(3), 393-408.

Prieler, J. A., & Raven, J. (2002). The measurement of change in groups and individuals, with particular reference to the value of gain scores: A new IRT-based methodology for the assessment of treatment effects and utilizing gain scores. *Horizons of Psychology*, *11*(3), 119-150.

Primi, R. (2001). Complexity of geometric inductive reasoning tasks: Contribution to the understanding of fluid intelligence. *Intelligence*, *30*, 41-70.

Primi, R., Eugénia Ferrao, E., & Almeida, L. (2010). Fluid intelligence as a predictor of learning: A longitudinal multilevel approach applied to math. *Learning and Individual Differences*, *20*, 445-451.

Rasch, G. (1961). On general laws and the meaning of measurement in psychology. In *Proceedings of the fourth berkeley symposium on mathematics* (p. 321-333). Statistics and Probability.

Raven, J. (1936). *Mental tests used in genetic studies: The performance of related individuals on tests mainly educative and mainly reproductive*. Unpublished master's thesis, University of London, London, UK.

Raven, J., Raven, J., & Court, J. H. (2004). *Manual for Raven's Progressive Matrices and*

Vocabulary Scales. San Antonio, Texas: Harcourt Assessment.

Resing, W. C. M. (1990). *Intelligentie en leerpotentieel. Een onderzoek naar het leerpotentieel van jonge leerlingen uit het basis- en speciaal onderwijs* [Intelligence and learning potential research on the learning potential of young children from mainstream and special education]. Lisse: Swets and Zeitlinger Inc.

Resing, W. C. M. (1993). Measuring inductive reasoning skills: The construction of a learning potential test. In J. H. M. Hamers, K. Sijtsma, & A. J. J. M. Ruijssenaars (Eds.), *Learning potential assessment: Theoretical, methodological and practical issues* (p. 219-241). Lisse: Swets and Zeitlinger Inc.

Resing, W. C. M. (1997). Learning potential assessment: the alternative for measuring intelligence? *Educational and Child Psychology*, *14*, 68-82.

Resing, W. C. M. (2000). Assessing the learning potential for inductive reasoning (LIR) in young children. In C. S. Lidz & J. G. Elliott (Eds.), *Dynamic assessment: Prevailing models and applications* (p. 229-262). Oxford: Elsevier Inc.

Resing, W. C. M., Bosma, T., & Stevenson, C. E. (2012). Dynamic testing: Measuring inductive reasoning in children with developmental disabilities and mild cognitive impairments. *Journal of Cognitive Education and Psychology*, *11*(2), 159-178.

Resing, W. C. M., & Elliott, J. G. (2011). Dynamic testing with tangible electronics: Measuring children's change in strategy use with a series completion task. *British Journal of Educational Psychology*, *81*(4), 579-605.

Resing, W. C. M., Elliott, J. G., & Grigorenko, E. (2012). Dynamic testing & assessment. In N. Seel (Ed.), *Encyclopedia of the sciences of learning.* New York: Springer Verlag.

Resing, W. C. M., & Roth-Van Der Werf, G. J. M. (2003). Teaching children to think inductively: Looking through the theoretical mirror. *Educational and Child Psychology*, *20*, 52-63.

Resing, W. C. M., Steijn, W. M. P., Xenidou-Dervou, I., Stevenson, C. E., & Elliott, J. G. (2011). Computerized dynamic testing: A study of the potential of an approach using sensor technology. *Journal of Cognitive Education and Psychology*, *10*(2), 178-194.

Resing, W. C. M., Tunteler, E., De Jong, F., & Bosma, T. (2009). Dynamic testing in indigenous and ethnic minority children. *Learning and Individual Differences*, *19*(4), 445-450.

Resing, W. C. M., Xenidou-Dervou, I., Steijn, W. M. P., & Elliott, J. G. (2012). A 'picture' of children's potential for learning: Looking into strategy changes and working memory by dynamic testing. *Learning and Individual Differences*, *22*(1), 144-150.

Richland, L. E., Morrison, R. G., & Holyoak, K. J. (2006). Children's development of analogical reasoning: Insights from scene analogy. *Journal of Experimental Child Psychology*, *94*, 249-273.

Rijmen, F., & De Boeck, P. A. L. (2001). Propositional reasoning: the differential contribution of "rules" to the difficulty of complex reasoning problems. *Memory and Cognition*, *29*(1), 165-175.

Rittle-Johnson, B. (2006). Promoting transfer: Effects of self-explanation and direct instruction. *Child Development*, *77*, 1-15.

Rizopoulos, D. (2006). `ltm`: An R package for latent variable modeling and item response analysis. *Journal of Statistical Software*, *17*(5), 1-25.

Roth-Van Der Werf, G. J. M., Resing, W. C. M., & Slenders, P. A. C. (2002). Task similarity and transfer of an inductive reasoning training. *Contemporary Educational Psychology*, *27*(2), 296-325.

Rutland, A., & Campbell, R. (1995). The validity of dynamic assessment methods for children with learning difficulties and nondisabled children. *Journal of Cognitive Education*, *5*, 81-94.

Samenjima, F. (1997). Graded response model. In W. J. Van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (p. 85-100). New York: Springer.

Schraw, G. (1998). Promoting general metacognitive awareness. *Instructional Science*, *26*, 113-125.

Siegler, R. S. (2002). Microgenetic studies of self-explanation. In N. Granott & J. Parziale (Eds.), *Microdevelopment: Transition processes in development and learning* (p. 31-58). Cambridge, UK: Cambridge University Press.

Siegler, R. S. (2006). Microgenetic analyses of learning. In D. Kuhn & R. Siegler (Eds.), *Handbook of child psychology: Cognition, perception, and language* (6th ed., Vol. 2, p. 464-510). New York: Wiley and Sons. (W. Damon and R. M. Lerner (Series Editors))

Siegler, R. S., & Svetina, M. (2002). A microgenetic/cross-sectional study of matrix completion: comparing short-term and long-term change. *Child Development*, *73*(3), 793-809.

Sittner Bridges, M., & Catts, H. (2011). The dynamic screening of a phonological awareness to predict risk for reading disabilities in kindergarten children. *Journal of Learning Disabilities*, *44*(4), 330-338.

Speece, D. L., Cooper, D. H., & Kibler, J. M. (1990). Dynamic assessment, individual differences, and academic achievement. *Learning and Individual Differences*, *2*, 113-127.

Stanovich, K. E., Cunningham, A. E., & Freeman, D. J. (1984). Intelligence, cognitive skills, and early reading progress. *Reading Research Quarterly*, *19*, 278-303.

Sternberg, R. J. (1977). Component processes in analogical reasoning. *Psychological Review*, *84*(4), 353-378.

Sternberg, R. J. (2004). Culture and intelligence. *American Psychologist*, *59*(5), 325-338.

Sternberg, R. J., & Gardner, M. K. (1983). Unities in inductive reasoning. *Journal of Experimental Psychology: General*, *112*(1), 80-116.

Sternberg, R. J., & Grigorenko, E. L. (2002). *Dynamic testing*. New York, United States of America: Cambridge University Press.

Sternberg, R. J., Grigorenko, E. L., & Bundy, D. A. (2001). The predictive value of IQ. *Merrill-Palmer Quarterly*, *47*(1), 1-41.

Sternberg, R. J., Grigorenko, E. L., Ngorosho, D., Tantufuye, E., Mbised, A., Nokes, C., et al. (2002). Assessing intellectual potential in rural Tanzanian school children. *Intelligence*, *30*, 141-162.

Sternberg, R. J., Grigorenko, E. L., Ngorosho, D., Tantufuye, E., Mbised, A., Nokes, C., et al. (2007). *Dynamic instruction for and assessment of developing expertise in four ethnic groups*. University of Connecticut: Storrs, CT: National Research Center on the Gifted and Talented.

Sternberg, R. J., & Kaufmann, S. B. (Eds.). (2011). *The cambridge handbook of intelligence*. New York, New York: Cambridge University Press.

Sternberg, R. J., & Rifkin, B. (1979). The development of analogical reasoning processes. *Journal of Experimental Child Psychology*, *27*, 195-232.

Stevenson, C. E., Heiser, W. J., & Resing, W. C. M. (submitted 2011a). Dynamic testing of ethnic minority children's potential for learning to solve analogies.

Stevenson, C. E., Heiser, W. J., & Resing, W. C. M. (submitted 2011b). Prompting learning and transfer of analogical reasoning: Is working memory a piece of the puzzle?

Stevenson, C. E., Heiser, W. J., & Resing, W. C. M. (submitted 2012b). Dynamic measures of analogical reasoning predict children's math and reading achievement.

Stevenson, C. E., Heiser, W. J., & Resing, W. C. M. (under review). Dynamic testing of analogical reasoning in 5-6 year olds: multiple-choice versus constructed-

response training.

Stevenson, C. E., Hickendorff, M., Heiser, W. J., Resing, W. C. M., & De Boeck, P. A. L. (submitted 2012a). Explanatory item response modeling of children's change on a dynamic test of analogical reasoning.

Stevenson, C. E., Resing, W. C. M., & Froma, M. N. (2009). Analogical reasoning skill acquisition with self-explanation in 7-8 year olds: Does feedback help? *Educational and Child Psychology*, *26*(3), 6-17.

Stevenson, C. E., Touw, K. W. J., & Resing, W. C. M. (2011). Computer or paper analogy puzzles: Does assessment mode influence young children's strategy development? *Educational and Child Psychology*, *28*(2), 67-84.

Süb, M., Oberauer, K., Wittmann, W. W., Wilhelm, O., & Schulze, R. (2002). Working-memory capacity explains reasoning ability - and a little bit more. *Intelligence*, *30*, 261-288.

Swanson, H. L. (1994). The role of working memory and dynamic assessment in the classification of children with learning disabilities. *Learning Disabilities Research and Practice*, *9*, 190-202.

Swanson, H. L. (2008). Working memory and intelligence in children: What develops? *Journal of Educational Psychology*, *100*, 581-602.

Swanson, H. L. (2011a). Does the dynamic testing of working memory predict growth in non-word fluency and vocabulary in children with reading disabilities? *Journal of Cognitive Education and Psychology*, *9*(2), 139-165.

Swanson, H. L. (2011b). Dynamic testing, working memory and reading comprehension growth in children with reading disabilities. *Journal of Learning Disabilities*, *44*(4), 358-371.

Swanson, H. L., & Lussier, C. M. (2001). A selective synthesis of the experimental literature on dynamic assessment. *Review of Educational Research*, *71*(2), 321-363.

Taub, G. E., Floyd, R. G., Keith, T. Z., & McGrew, T. Z. (2008). Effects of general

and broad cognitive abilities on mathematics achievement. *School Psychology Quarterly*, *23*, 187-198.

Te Nijenhuis, J., & Van Der Vlier, H. (2001). Group differences in mean intelligence for the dutch and third world immigrants. *Journal of Biosocial Science*, *33*, 469-475.

Thatcher-Kantor, P., Wagner, R. K., Torgensen, J., & Rashotte, C. A. (2011). Comparing two forms of dynamic assessment and traditional assessment of preschool phonological awareness. *Journal of Learning Disabilities*, *44*(4), 313-321.

Thibaut, J. P., French, R. M., & Vezneva, M. (2008). Analogy-making in children: The importance of processing constraints. In *Proceedings of the thirtieth annual cognitive science society conference* (p. 475-480).

Thibaut, J. P., French, R. M., & Vezneva, M. (2010). The development of analogy making in children: Cognitive load and executive functions. *Journal of Experimental Child Psychology*, *106*, 1-19.

Thorell, L. B., Lindqvist, S., Nutley, B., S Bohlin, G., & Klingberg, T. (2009). Training and transfer effects of executive functions in preschool children. *Developmental Science*, *12*(1), 106-113.

Tillman, C. M., Nyberg, L., & Bohlin, G. (2008). Working memory components and intelligence in children. *Intelligence*, *36*, 394-402.

Tissink, J., Hamers, J. H. M., & Van Luit, J. E. H. (1993). Learning potential tests with domain-general and domain-specific tasks. In J. H. M. Hamers, K. Sijtsma, & A. J. J. M. Ruijssenaars (Eds.), *Learning potential assessment: Theoretical, methodological and practical issues* (p. 243-266). Lisse: Swets and Zeitlinger Inc.

Tunteler, E., Pronk, C. M. E., & Resing, W. C. M. (2008). Inter- and intra-individual variability in the process of change in the use of analogical strategies to solve geometric tasks in children: A microgenetic analysis. *Learning and Individual Differences*, *18*(1), 44-60.

Tunteler, E., & Resing, W. C. M. (2002). Spontaneous analogical transfer in 4-

year-olds: A microgenetic study. *Journal of Experimental Child Psychology*, *83*(3), 1-19.

Tunteler, E., & Resing, W. C. M. (2007a). Change in spontaneous analogical transfer in young children: A microgenetic study. *Journal of Infant and Child Development*, *16*, 71-94.

Tunteler, E., & Resing, W. C. M. (2007b). Change in spontaneous analogical transfer in young children: a microgenetic study. *Infant and Child Development*, *16*(1), 71-94.

Tunteler, E., & Resing, W. C. M. (2007c). Effects of prior assistance on young children's unprompted analogical problem solving over time: A microgenetic study. *British Journal of Educational Psychology*, *77*(1), 43-68.

Tunteler, E., & Resing, W. C. M. (2010). The effects of self- and other-scaffolding on progression and variation in children's geometric analogy performance: A microgenetic research. *Journal of Cognitive Education and Psychology*, *9*(3), 251-272.

Tzuriel, D. (2001). *Dynamic assessment of young children*. New York: Kluwer Academic/Plenum Publishers.

Tzuriel, D., & Egozi, G. (2010). Gender differences in spatial ability of young children: The effects of training and processing strategies. *Child Development*, *81*(5), 1417-1430.

Tzuriel, D., & Kaufman, R. (1999). Mediated learning and cognitive modifiability: Dynamic assessment of young ethiopian immigrant children to israel. *Journal of Cross-cultural Psychology*, *30*(3), 359-380.

Vakil, E., Lifshitz, H., Tzuriel, D., Weiss, I., & Arzuoan, Y. (2010). Analogies solving by individuals with and without intellectual disability: Different cognitive patterns as indicated by eye-movements. *Research in Developmental Disabilities*, *32*, 846-856.

Van de Vijver, F. R. (2002). Inductive reasoning in Zambia, Turkey, and the

Netherlands: Establishing cross-cultural equivalence. *Intelligence*, *30*, 313-351.

Van de Vijver, F. R. (2008). On the meaning of cross-cultural differences in simple cognitive measures. *Educational Research and Evaluation*, *14*(3), 215-234.

Van de Vijver, F. R., & Poortinga, Y. H. (1997). Towards an integrated analysis of bias in cross-cultural assessment. *European Journal of Psychological Assessment*, *13*(1), 29-37.

Verguts, T., & De Boeck, P. A. L. (2000). A rasch model for detecting learning while solving an intelligence test. *Psychological Measurement*, *24*(2), 151-162.

Vigneau, F., Caissie, A. F., & Bors, D. A. (2006). Eye-movement analysis demonstrates strategic influences on intelligence. *Intelligence*, *34*, 261-272.

Vock, M., Prekel, F., & Holling, H. (2011). Mental abilities and school achievement: A test of mediation hypothesis. *Intelligence*, *39*(5), 357-369.

Von Davier, M., Xu, X., & Carstensen, C. (2010). Measuring growth in a longitudinal large-scale assessment with general latent variable modeling. *Psychometrika*, *76*(2), 318-336.

Vygostsky, L. S. (1978). *Mind in society: The development of higher psychological processes, (1938)* (M. Cole, V. John-Steiner, S. Scribner, & E. Sonberman, Eds.). Cambridge, MA: Harvard University Press.

Wang, Y., & Heffernan, N. T. (2011). The "assistance" model: Leveraging how many hints and attempts a student needs. In *Proceedings of the 24th International Florida Artificial Intelligence research society conference* (p. 549-554).

Wechsler, D. (2003). *Wechsler intelligence scale for children-fourth edition*. Administration and scoring manual. San Antonio, TX: Harcourt Assessment, Inc.

Wiedl, K. H., Kampling, V., Köning, I., Schrevels, E. M., & Waldorf, M. (2011, September). *Availability of a German version of the application of cognitive functions scale (ACFS) for the assessment of retarded children.*

Wiedl, K. H., Schöttke, H., Green, M. F., & Nuechterlein, K. H. (2004). Dynamic

testing in schizophrenia: Does training change the construct validity of a test? *Schizophrenia Bulletin*, *30*(4), 703-711.

Summary in Dutch (Samenvatting)

Onderzoek naar de cognitieve ontwikkeling van kinderen laat zien dat grote verschillen niet alleen optreden in wat kinderen al kunnen maar ook in *hoe* ze leren. Dynamisch testen is een methode om cognitieve vaardigheden in ontwikkeling – zoals bijvoorbeeld het redeneervermogen – te meten. Het gaat bij dynamisch testen niet alleen om wat een kind al weet, maar vooral om zijn of haar vermogen om te leren (Elliott, 2003; Sternberg & Grigorenko, 2002). Dit doel onderscheidt dynamische tests van conventionele, statische tests, zoals intelligentietests. Ondanks het feit dat statische tests veelvuldig gebruikt worden als er vragen zijn over de schoolprestaties van een kind, zijn ze bekritiseerd omdat ze vooral de huidige cognitieve vaardigheden en niet zozeer het potentieel van een kind in kaart brengen. Met dynamische tests kan informatie over het cognitief potentieel en instructiebehoefte van een kind verkregen worden en dit kan belangrijk zijn voor keuzes over onderwijs (Bosma & Resing, 2012). Dynamisch testen onderscheidt zich van conventionele testsituaties omdat er training wordt gegeven in aanvulling op een of meer statische testmomenten. Zo kan worden nagegaan hoe en in hoeverre het kind leert gedurende het hele traject van voormeting tot en met nameting (Elliott

et al., 2010).

De doelstelling van dit promotieonderzoek was om een dynamische test voor basisschoolleerlingen te ontwikkelen die het leervermogen van een kind op het gebied van analogisch redeneren in kaart brengt. Analogietaken zijn gekozen omdat deze, vanwege de complexe wijze waarop zulke taken opgelost dienen te worden, vaak gebruikt worden in intelligentietests en omdat het analogisch redeneervermogen relevant is voor het schoolse leren (Goswami, 1992). Een eerste streven was inzicht te krijgen in de factoren die een rol spelen in de grote variatie in leervermogen van kinderen. In de verschillende studies in dit proefschrift zijn twee factoren onderzocht: (1) vorm van training – zowel type instructies als type opgaven en (2) persoonskenmerken zoals etniciteit en werkgeheugen. Een tweede streven was te bepalen of leervermogen gemeten met deze test schoolse prestaties voorspelt.

In hoofdstuk 1 zijn de algemene uitgangspunten van de ontwikkelde dynamische test en de factoren die mogelijk van invloed zijn op de prestatie op deze test besproken. Dynamisch testen werd in dit proefschrift opgevat als een methode gericht op het in kaart brengen van het cognitief potentieel en het leerproces van een kind tijdens een testafname. De in dit proefschrift gebruikte dynamische test, AnimaLogica, bestond uit een voortoets, gevolgd door een korte training en een natoets. De *voortoets* geeft een indicatie van het *huidige analogisch redeneervermogen* – een meting waarbij geen hulp of feedback wordt geboden (Resing, 1997). De voortoets wordt gevolgd door twee *trainingen* waarin het kind volgens de 'graduated prompts'-methode getraind wordt. 'Graduated prompting' is een stapsgewijze trainingsmethode waarbij volgens een hiërarchisch principe zo weinig mogelijk hulp wordt geboden om het kind zo zelfstandig mogelijk de taak te laten oplossen (bijv. Campione & Brown, 1987; Resing, 1993; Resing & Elliott, 2011). Eerst werden algemene, metacognitieve instructies gegeven die het plannen stimuleerden of de

aandacht op de taak richtten. Daarna werd specifiekere hulp gegeven waardoor het kind steeds meer inzicht kreeg in hoe de taak opgelost diende te worden. Als deze stappen er nog niet toe leidden dat het kind de juiste oplossing vond, dan maakte de trainer de opgave samen met het kind. Het achterliggende idee is dat het kind alleen hulp krijgt als dat nodig is, zodat de instructiebehoefte van het kind gemeten kan worden. De *hoeveelheid* benodigde instructie tijdens de training geeft een indicatie van het leervermogen van het kind. De *typen* instructies die tot zelfstandige oplossingen hebben geleid geven een indicatie van welke instructies mogelijk ook op school effectief zouden kunnen zijn. De trainingssessies worden gevolgd door een *natoets*, die net zoals de voortoets zonder hulp of feedback werd afgenomen. De natoets geeft het *potentieel vermogen* aan – wat een kind mogelijk zou kunnen met geindividualiseerde instructie. Het analyseren van de *uitleg* van het kind en welke *strategieën* hij of zij heeft toegepast geeft informatie over het leerproces – oftewel hoe het kind geleerd heeft tijdens de dynamische test (bijv. Resing et al., 2009). Het vermogen om de geleerde kennis spontaan op een nieuw probleem toe te passen, zogeheten *transfer*, geeft aan in hoeverre het kind na een korte interventie begrijpt wat analogisch redeneren is (bijv. Campione et al., 1985; Ferrara et al., 1986; Resing, 1990).

Twee facetten die het regelmatig gebruik van dynamische testen bemoeilijken zijn in AnimaLogica meegenomen: (1) de duur van een dynamische test en (2) de manier waarop verandering in prestatie wordt gemeten (Grigorenko & Sternberg, 1998). De afname van AnimaLogica is aanzienlijk korter dan bij eerdere dynamische tests het geval was – ongeveer 80 minuten – hetgeen overeenkomt met de duur van andere cognitieve tests. Dit is o.a. bereikt door het verkorten van de trainingstijd en gebruik te maken van een taak die gemakkelijk op de computer afgenomen kon worden (zie Stevenson et al., 2011 voor een bespreking van papieren versus computerafname). De psychometrische kwaliteit van dynamische tests is vaak

onduidelijk of wordt als onvoldoende beschouwd. De reden hiervoor is dat de klassieke wijze van het meten van de *mate van verandering* – door simpelweg het aantal goede oplossingen op de voortoets en de natoets te vergelijken – door psychometrici als onbetrouwbaar wordt beschouwd (De Bock, 1976). Een bijkomend probleem is dat een verschil van bijvoorbeeld vier juiste antwoorden een andere waarde kan hebben voor een kind dat oorspronkelijk twaalf opgaven goed had of een kind dat maar één opgave van de twintig goed had (Embretson, 1991b). Welk kind heeft meer geleerd? Door de problemen met betrouwbaarheid kan de mate van verandering op basis van ruwe scores beter niet gebruikt worden in het dynamisch testonderzoek (Resing, Elliott, & Grigorenko, 2012). Toch kan de mate van verandering mogelijk waardevolle informatie opleveren over het leervermogen als deze op een andere wijze – met behulp van item-respons theorie – wordt berekend (Embretson & Reise, 2000). Het hoofddoel van de (nog gaande) ontwikkeling van AnimaLogica was de instructiebehoefte en het potentieel van een kind te meten terwijl rekening werd gehouden met psychometrische standaarden en een korte, simpele afname.

Kinderen vertonen grote verschillen in zowel instructiebehoefte als mate van verandering in hun prestaties op een dynamische test (bijv. Resing et al., 2009). Met dynamisch testen wordt getracht deze verschillen te meten. Het doel van dynamisch testen is dus niet om blijvende verandering aan te brengen, maar om het leerpotentieel en het leerproces in kaart te brengen (Resing, Elliott, & Grigorenko, 2012). Het gemeten leervermogen wordt echter beïnvloed door factoren als de *vorm* van training en ook door *kenmerken van het kind*.

In hoofdstuk 2 is aandacht besteed aan de rol die de vorm van de opgaven speelt bij het verkrijgen van inzichten in het leervermogen van een kind. Kinderen uit groep 2 kregen ofwel 'graduated prompts'-training met meerkeuzevragen, ofwel 'graduated prompts'-training met open vragen waarbij het antwoord

geconstrueerd moest worden, ofwel geen training. De twee groepen die 'graduated prompts'-training kregen lieten na training meer progressie in analogisch redeneren zien dan de controlegroep. Dit schetst het beeld dat de 'graduated prompts'-training gemiddeld gezien- een effectieve manier is om het analogisch redeneren van vijf- en zesjarigen te stimuleren. Er was geen verschil in de mate van vooruitgang van voortoets naar natoets tussen de twee trainingsgroepen, maar de 'antwoordconstructiegroep' kon gemiddeld gezien wel betere uitleg geven van hun antwoorden en lieten een ander strategiegebruik zien dan de kinderen in de 'meerkeuzegroep'. Als antwoord op de hoofdvraag of trainen met meerkeuze- dan wel met antwoordconstructieopgaven het meest geschikt zou zijn voor dynamisch testonderzoek werd geconcludeerd dat antwoordconstructie een specifieker inzicht geeft in het redeneerproces van de kinderen. Daarom werd gekozen voor antwoordconstructieopgaven in de dynamische test afgenomen in het vervolgonderzoek.

In hoofdstuk 3 werd onderzocht of de ontwikkelde dynamische test geschikt is voor zowel autochtone als allochtone leerlingen. Bij traditioneel afgenomen intelligentietests zijn kinderen van de dominante cultuur over het algemeen in het voordeel (bijv. Van de Vijver, 2002). Dit kan bijvoorbeeld komen door verschillen in taalvaardigheid of verschillen in ervaring met soortgelijke opgaven of testsituaties. Deze problemen kunnen ertoe leiden dat er een vertekend beeld ontstaat van de huidige vermogens en het leerpotentieel van etnische minderheden vergeleken met die van hun autochtone leeftijdsgenoten (Sternberg et al., 2002). Item-respons theorie werd toegepast voor het meten van vooruitgang en werd rekening gehouden met onder andere de persoonsfactor werkgeheugen. In dit onderzoek waren autochtone en allochtone leerlingen verdeeld over drie groepen: 'graduated prompts'-training, zelfstandig oefenen met de opgaven of geen training (controle). Er werden geen verschillen gevonden tussen autochtone en allochtonen leerlingen in de

'graduated prompts'-groep in mate van vooruitgang, strategiegebruik, behoefte aan instructie of uitleg van hun oplossingen tijdens de training. Hieruit werd geconcludeerd dat de dynamische test ingezet kan worden voor het meten van leervermogen bij cultureel-diverse schoolpopulaties. Werkgeheugen bleek niet tussen beide leerlinggroepen te verschillen, en was bij beide groepen gerelateerd aan het analogisch redeneervermogen.

In hoofdstuk 4 werd de samenhang tussen twee vormen van werkgeheugen, het verbale en visuo-spatiele werkgeheugen, en de prestaties op de dynamische test onderzocht. De focus hierbij lag op transfer – oftewel het spontaan kunnen toepassen van hetgeen tijdens de trainingen is geleerd op andere, gerelateerde opgaven (Jacobs & Vandeventer, 1971). Twee groepen leerlingen, verdeeld over een 'graduated prompts'-trainingsgroep en een controle groep dat oefende met dezelfde opgaven, participeerden in het onderzoek. Bij de voormeting en nameting werden naast de analogieën met dierenfiguren ook twee andere redeneertaken, plus een 'reversal' taak afgenomen, waarbij het kind opgaven dient te ontwerpen voor de trainer en uitleg moet geven hoe de taken opgelost kunnen worden (Bosma & Resing, 2006). De kinderen die beter presteerden op de voormeting bleken over het algemeen een efficiënter werkgeheugen te hebben. De *mate* van vooruitgang bleek echter geen verband te houden met werkgeheugen. De prestaties op de transfertaken bij de nameting waren enigszins gerelateerd aan de prestaties op de voormeting. Bij transfer speelde het werkgeheugen wederom geen rol. Redeneervermogen en werkgeheugen zijn twee constructen die vaak gemeten worden als een schoolpsycholoog inzicht wil krijgen in de cognitieve capaciteiten van een kind. Deze constructen bleken weinig samenhang te vertonen met maten voor leervermogen en transfer. Dit betekent dat leervermogen en transfer belangrijk zouden kunnen zijn bij het in kaart brengen van het cognitief potentieel van een kind.

In hoofdstuk 5 werd dieper ingegaan op de meting van de mate van vooruitgang tussen de voormeting en nameting en de samenhang hiervan met het werkgeheugen. Het meten van verandering wordt door psychometrici onbetrouwbaar geacht wanneer er sprake is van verschillen in percentage goed tussen de voor- en nameting (Lord, 1963). Item-respons theorie biedt mogelijkheden om de mate van vooruitgang op betrouwbare wijze te meten (Embretson & Reise, 2000). Item-respons theorie werd in de studie in dit hoofdstuk gebruikt om niet alleen mate van vooruitgang te meten maar ook binnen eenzelfde statistisch model de verschillen in leervermogen tussen kinderen te verklaren aan de hand van enerzijds het type training dat werd gegeven plus anderzijds persoonskenmerken zoals werkgeheugen. Basisschoolleerlingen uit groepen twee, drie en vier waren verdeeld over twee trainingcondities: 'graduated prompts' en feedback. Bij de feedbacktraining kreeg het kind net zoals in de 'graduated prompts' training vijf kansen om het goede antwoord te construeren. In tegenstelling tot de 'graduated prompts' getrainde kinderen kregen ze geen instructies over hoe ze dat moesten doen, maar kregen ze alleen te horen of hun antwoord goed of fout was. De mate van vooruitgang van de kinderen in de feedbackconditie bleek minder sterk te zijn dan die van de kinderen in de 'graduated prompts'-groep. In beide gevallen was het werkgeheugen geen verklarende factor voor de mate van vooruitgang. De prestaties op de voortoets hingen samen met leeftijd, maar leeftijd was geen verklarende factor van de individuele verschillen in vooruitgang. Er was enig verband te zien tussen de prestaties van een kind op de voormeting en zijn mate van vooruitgang bij de natoets. Dit gaf echter geen volledig beeld van het leervermogen. Wel bleek dat kinderen die hoge scores behaalden op landelijke rekentoetsen ook beter presteerden op de voormeting en ook meer vooruitgang lieten zien tijdens het dynamisch testen. Dit ondersteunt eerdere conclusies dat de mate van vooruitgang mogelijk een belangrijk construct vormt bij het meten van het cognitief potentieel van een kind

(bijv. Embretson & Prenovorst, 2000).

De voorspellende waarde van de mate van vooruitgang gemeten met AnimaLogica op schoolprestaties werd onderzocht in hoofdstuk 6. Conventionele tests, zoals een intelligentietest, hebben enige voorspellende waarde ten aanzien van toekomstige schoolprestaties (Sternberg et al., 2001). Dynamische testuitkomsten lijken van toegevoegde waarde te zijn ten aanzien van deze voorspelling (Caffrey et al., 2008). Het is echter niet duidelijk welk aspect van de metingen het meeste bijdraagt aan de voorspelling: de instructiebehoefte tijdens de training, de mate van vooruitgang of het transfervermogen. Dit onderzoek bouwt voort op eerdere onderzoek (bijv. Beckmann, 2006; Resing, 1993), maar voegde ook drie aspecten toe: (1) de opzet was longitudinaal, (2) de testgroep bestond uit reguliere basisschoolkinderen, en (3) de voorspelling was op nationaal genormeerde toetsen toegepast. Kinderen uit groep drie van de basisschool werden dynamisch getest met de 'graduated prompts'-methode. Van elk kind zijn de prestaties voor rekenen en lezen, afkomstig uit de gegevens van het leerlingvolgsysteem, verzameld op drie momenten: 3 weken voor het dynamisch testen, 5 maanden na het dynamisch testen en 1 jaar later. In deze studie zijn conventionele en dynamische testgegevens vergeleken bij het voorspellen van de scores op rekenen en lezen. De prestatie op de voortoets werd beschouwd als een conventionele meting van het analogisch redeneren. De mate van vooruitgang, instructiebehoefte tijdens training en prestaties op de 'reversal' transfertaak waren de dynamische metingen. De mate van vooruitgang bleek de beste voorspeller van de scores van de kinderen in zowel rekenen als lezen. Deze vondst gaf verdere ondersteuning voor de hypothese dat dynamische testgegevens van toegevoegde waarde kunnen zijn bij het in kaart brengen van het cognitief potentieel van een kind.

Ten slotte werd in hoofdstuk 7 geconcludeerd dat hoewel *huidige prestaties* op analogietaken beïnvloed worden door persoonskenmerken zoals leeftijd,

werkgeheugen en etniciteit, deze factoren het *leervermogen* gemeten met een dynamische test niet verklaren. Het type training dat gegeven wordt heeft echter wel invloed op de mate van vooruitgang van de voormeting naar de nameting. Uit het onderzoek gepresenteerd in dit proefschrift blijkt dat 'graduated prompting' tot grotere vooruitgang leidt dan feedbacktraining of zelfstandig oefenen. Ook de vorm van de opgaven speelt een rol waarbij antwoord-constructie of zelfs opgavencreatie, zoals in de 'reversal' taak, een meer volledig beeld van het leerpotentieel geven.

De algemene conclusie in dit proefschrift is dat uitkomsten op een dynamische test met een 'graduated prompts'-training van toegevoegde waarde kunnen zijn wanneer onderwijzers zich afvragen wat het leerpotentieel van een kind is. Het geeft mogelijk ook een eerlijker beeld van het leervermogen van allochtone leerlingen die op conventionele tests in het nadeel zijn. Er blijft echter grote variabiliteit in de prestaties en vooruitgang van kinderen op analogisch redeneertaken. De individuele verschillen in leervermogen gemeten met deze dynamische test zijn niet eenduidig of gemakkelijk te verklaren uit persoonskenmerken of trainingsvorm, maar ze geven wel informatie die toekomstige schoolprestaties in rekenen en lezen kan helpen voorspellen. Een dynamische test zou ingezet kunnen worden om na te gaan welke kinderen meer potentieel hebben dan wat op dit moment uit de schoolprestaties blijkt. Ook zou een dynamische test vroegtijdig kunnen signaleren welke kinderen dreigen achter te lopen. Hierbij zou instructiebehoefte en transfervermogen waardevolle informatie kunnen bieden zodat onderwijzers een passende interventie kunnen ontwikkelen dat een kind helpt zijn of haar cognitief potentieel optimaal te benutten.

Propositions

I. Performance change during dynamic testing is an important construct in the assessment of learning and cognitive potential. (Chapter 7, this thesis)

II. Item response theory is an appropriate method for measuring individual differences in change in dynamic tests as it provides a good basis for the latent scaling of gain scores. (Chapter 5, this thesis)

III. Analogy item format influences children's performance and item effects should be taken into consideration when measuring potential for learning to solve analogies. (Chapter 2, this thesis)

IV. Dynamic testing of analogical reasoning with AnimaLogica has potential as a multicultural dynamic assessment instrument. (Chapter 3, this thesis)

V. Children with untapped cognitive potential are more often present in low functioning groups and are likely to be overlooked if they are judged based on conventional, static reasoning tests. (Chapter 5, this thesis)

VI. If teachers knew for which children the saying "little help can go a long way" rings true, realizing potential in the classroom may be more manageable.

VII. Cognitive potential is a like the stretchability of a metal spring: how far ability can be stretched and how much effort it takes to reach this maximum reveals what one is truly capable of.

VIII. Transfer of knowledge to new situations is the aim of all schooling, but difficult to induce in an experimental setting. (Opfer & Thompson, 2008)

IX. "Young children and other animals" reason by association, whereas humans can induce new rules. (Kendler & Kendler, 1962)

X. Educational technology has potential for enhancing *all* children's learning - especially if it's freely available, language independent and adapts to provide stimuli and feedback that meet the Goldilocks requirement of being *just right*.

XI. In children's learning, micro-development may reflect macro-development under similar conditions. (Häckel's Law)

Curriculum vitae

Claire Stevenson (1976) was born in Baton Rouge, Louisiana in the USA. She attended elementary and middle school first in Monroe, Louisiana then in Kansas City, Missouri. She completed the International Baccalaureate program at Maartenscollege international high school in Haren in the Netherlands and obtained her doctorandus (Master's) degree in Psychology with honors in 2001 at Leiden University. Her master's thesis focused on the development of kindergartner's understanding of multiplication under the supervision of Dr. Anke Blöte. After completing an intense and rewarding internship at the Monroe City Schools School Office of Pupil Appraisal under the supervision of Robin Cohenour, school psychologist, Claire wanted to continue in the field of educational psychology with a focus on the measurement and training of children's learning and thinking processes but on a larger scale. This sparked an interest in educational technology for which Claire began studying Computer Science at Leiden University in 2002 and soon after started a professional career as a software developer. However, children's cognitive development was not the main focus in this career path. Therefore she welcomed the opportunity to conduct research on the (computerized) assessment of children's learning potential with Prof. dr. Wilma Resing. For this project she returned to the Department of Developmental and Educational Psychology of Leiden University in 2007 and worked towards a PhD with Prof. Dr. Wilma Resing and Prof. Dr. Willem Heiser. Claire is currently working as a postdoctoral researcher at the Educational Neuroscience department of Prof. dr. Jelle Jolles at the VU University Amsterdam.