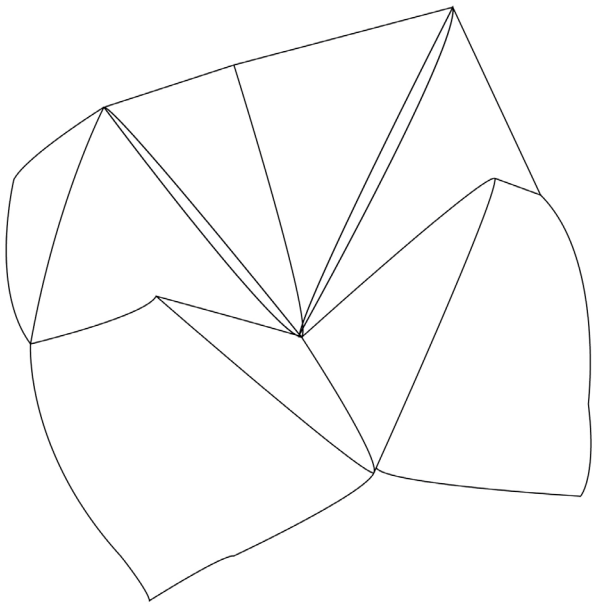
Dynamic testing and excellence: unfolding potential

Vogelaar, B.

CHAPTER 4

Dynamic testing of gifted and average-ability children's analogy problem-solving: Does executive functioning play a role?



Bart Vogelaar
Merel Bakker
Lianne Hoogeveen
Wilma C. M. Resing

Vogelaar, B., Bakker, M., Hoogeveen, L., & Resing, W. C. M. (submitted). Dynamic testing of gifted and average-ability children's analogy problem-solving: Does executive functioning play a role? *[Psychology in the Schools]*

Abstract

In this study, dynamic testing principles were applied to examine progression of analogy problem-solving, the roles that cognitive flexibility and metacognition play in children's progression as well as training benefits, and instructional needs of 7-8 year old gifted and average-ability children. Utilizing a pre-test-training-post-test control group design, participants were split in four subgroups: gifted dynamic testing (n=22), gifted unguided practice (n=23), average-ability dynamic testing (n=31) and average-ability unguided practice (n=37). Results revealed that dynamic testing led to more advanced progression than unguided practice, and that gifted and average-ability children showed equivalent progression lines and instructional needs. For children in both ability categories, cognitive flexibility was not found to be related to progression in analogy problem-solving or training benefits. In addition, metacognition was revealed to be associated with training benefits. Implications for educational practice were provided in the discussion.

4.1. Introduction

It has been proposed that cognitive abilities play an important role in children's school performance. Both intelligence (Balboni, Naglieri, & Cubelli, 2010; Roth, Becker, Romeyke, Schäfer, Domnick et al., 2015), and executive functions (e.g., Blair & Diamond, 2008; Monette, Bigras, & Guay, 2011; Viterbori, Usai, Traverso, & De Franchis, 2015) have been shown to predict school success. When a child is considered to be gifted in an educational context, this is often based on the results of an assessment procedure, including conventional, static testing of intelligence, or school aptitude (Kline, 2001). These tests, however, have been shown not to be advantageous for all children, and do not unveil information about psychological processes involved in learning (e.g., Grigorenko, 2009). As conventional tests, for a large part, rely on past learning experiences (Elliott, Grigorenko, & Resing, 2010), children who have had less than favorable learning experiences, have been documented to underperform on these tests (Robinson-Zañartu & Carlson, 2013). Dynamic tests, in contrast, are much more focused on a child's potential for learning, rather than on past learning experiences (Sternberg & Grigorenko, 2002). As in these tests, feedback and/or instruction are integrated into the testing procedure (Elliott, 2003), they allow for examining to what extent children show improvement in performance after an intervention, and whether other cognitive factors, such as executive functions, play a role in learning. In the current study, dynamic testing principles were applied to investigate to what extent two aspects of executive functioning, cognitive flexibility and metacognition, would be related to static or dynamic progression in analogy problem-solving of gifted and average-ability children.

Dynamic testing

Rather than measuring the knowledge or skills a child has already mastered, dynamic testing focuses on what a child would achieve in a short time-frame, and this assessment procedure is therefore expected to provide a more complete picture of a child's potential for learning (Elliott, 2003). The pre-test-training-post-test design (Sternberg & Grigorenko, 2002) is a frequently used application of dynamic testing that allows for structured measuring of a child's learning progression. The graduated prompts technique (e.g., Campione & Brown, 1987) has been used successfully as a training intervention in combination with said design. In this training approach, children are provided with structured prompts each time they make a mistake in problem solving. In the current study, prompts were tailored to each individual problem to be solved, and became more specific gradually, ranging from metacognitive to cognitive prompts and modelling (Resing & Elliott, 2011).

Similar to static test scores, dynamic testing outcomes have shown that there are many individual differences between children; both in terms of the instruction they require in order to show learning progression, as well as in terms of the level of progression they show after training (e.g., Resing, 2013, Sternberg & Grigorenko, 2002). Dynamic testing of children who have strong cognitive capacities, nevertheless, seems an area researched less intensively. Previous research indicates that gifted children not only have a cognitive advantage, but, more specifically, have a more extensive zone of proximal development, learn new skills faster, and are better at generalizing newly acquired knowledge (Calero, García-Martín, & Robles, 2011; Kanevsky, 2000). The potential role of executive functioning in dynamic testing of this group of children has, however, not yet been examined abundantly.

Dynamic tests frequently utilize inductive reasoning tasks (e.g., Ferrara, Brown, & Campione, 1986; Resing, 2000). Inductive reasoning is believed to play a central role in intelligence (Klauer & Phye, 2008), and is said to be of crucial importance with regard to acquiring and applying knowledge (Goswami, 2012) and solving problems (Richland & Burchinal, 2012).

Executive functioning

The graduated prompts technique employed in the current study included prompts activating different aspects of executive functioning, for example in relation to self-regulation and monitoring of the problem-solving process. Executive functions comprise a number of complex cognitive processes enabling conscious control of thought and action (Monette et al., 2011) that are critical to purposeful, goal-directed behavior (Arffa, 2007). They are seen as the cognitive component of self-regulation (Calkins & Marcovitch, 2010). Metacognition, a specific aspect of executive functioning, is usually described as consisting of self-reflective cognitive processes (Schneider, 2010), divided into two dimensions: knowledge, and regulation of cognitive activity (Moses & Baird, 1999), and is asserted to play an important role in developing new expertise (e.g., Sternberg, 1998).

In addition, it has been argued that flexibility in applying newly learned skills and knowledge can be seen as an important aspect of intellectual and cognitive functioning (e.g., Resing, 2013). Cognitive flexibility is said to include the ability to change perspectives spatially, or interpersonally, and being sufficiently flexible to adjust thinking to changing demands. Further, it is seen as a key component of the ability to think outside the box, and shares many characteristics with creativity, task and set switching (Diamond, 2013).

Executive functioning has been found to be related to cognition (e.g., Ardila, Pineda & Rosselli, 2000). Studies investigating the relationship of executive

functioning in a dynamic testing context, in particular with gifted children, however, are few, with most studies focusing on the role of working memory (e.g. Resing, Xenidou-Dervou, Steijn, & Elliott, 2012; Stevenson, Bergwerff, Heiser, & Resing, 2014; Stevenson, Heiser, & Resing, 2013; Swanson, 2006, 2010, 2011).

The current study

The current study utilized a dynamic test for analogical problem solving, a subtype of inductive reasoning, employing graduated prompts techniques. As studies have shown that analogical reasoning develops greatly in 7-8 year old children (e.g., Richland, Morrison, & Holyoak, 2006; Tunteler & Resing, 2007), children of this age group participated in this study. Our main research aim was to provide more insight into the potential benefits of dynamic testing of gifted children. More specifically, we focused on the roles that ability, cognitive flexibility and metacognition play in repeatedly measured static versus dynamic progression in solving analogies.

Our first cluster of research questions addressed children's progression in solving analogies from pre-test to post-test. Based on previous research into progression of unprompted solving of analogy problems amongst young children (e.g. Tunteler & Resing, 2007; Tunteler, Pronk, & Resing, 2008), we expected a significant main effect of time. We hypothesized (1a) that both unguided practice, and dynamic testing would lead to progression in solving analogies from session to session. More importantly, we expected a significant interaction of time x condition, hypothesizing (1b) that children in the dynamic testing condition would show more progression from pre-test, before training, to post-test, after training (e.g., Resing & Elliott, 2011; Stevenson et al., 2013). As our study focused on potential differences between gifted and average-ability children in relation to their progression, we expected a significant interaction between time and ability. Gifted children were reported to have a more extensive zone of proximal development (e.g., Calero et al., 2011; Kanevsky, 2000), therefore we hypothesized (1c) that gifted children would show more progression after unguided practice experiences than their average-ability peers. We also expected a significant interaction of time x condition x ability, indicating that gifted children would show more progression after training than their average-ability peers (1d).

Our second cluster of research questions concerned the association between executive functioning and children's progression from pre-test to post-test. We expected a significant interaction between time and cognitive flexibility. Considering that flexibility in applying skills and knowledge is suggested to be important for learning and applying new knowledge (e.g., Resing, 2013), we hypothesized (2a) that children with higher levels of cognitive flexibility would

show more progression in solving analogies than their peers with lower levels of cognitive flexibility. We also expected an interaction between time, condition, and cognitive flexibility, (2b) hypothesizing that children with higher levels of cognitive flexibility would benefit more from dynamic training than those with lower levels. Furthermore, a significant interaction between time, condition, ability and cognitive flexibility was expected. Building on empirical studies in which high-ability children were found to have an advantage in executive functioning (e.g., Arffa, 2007), we hypothesized (2c) that the progression paths of gifted children with higher levels of cognitive flexibility would be steeper than those of their average-ability peers with similar levels of cognitive flexibility.

Moreover, as self-regulating, metacognitive skills were found to play a significant role in learning (e.g., Campione, Brown, & Ferrara, 1982; Sternberg, 1998), we expected an interaction between time and metacognition, hypothesizing (3a) that children with higher levels of metacognition would show more progression in solving analogies than their peers with lower levels of metacognition. We also expected a significant interaction between time, metacognition and condition, and hypothesized (3b) that children with higher levels of metacognition would benefit more from training than their age-mates with lower levels of metacognition. Finally, a significant interaction was expected between time, condition, ability and metacognition. Taking into account that high-ability children were found to have an advantage in self-regulation (e.g., Calero, García-Martín, Jiménez, Kazén, & Araque, 2007), we hypothesized (3c) that the progression paths after training of the gifted children who have higher levels of metacognition would be steeper than their average-ability peers with similar levels of metacognition.

Our last research question focused on more closely to what extent gifted and average-ability children have different instructional needs, as measured by the number and the type of prompts required during training. As high-ability children were found to be more responsive to feedback (Kanevsky & Geake, 2004), and were found to have an advantage in self-regulation (e.g., Calero et al., 2007), we expected that gifted children's instructional needs during dynamic training would be significantly different from their average-ability peers. We hypothesized that gifted children would (4a) need both less metacognitive and (4b) less cognitive prompts than their average-ability peers.

4.2. Method

Participants

In the current study, 113 children, 54 boys and 59 girls, participated, ranging in age from 7;1 to 8;9 years (*M*=7.90). The average-ability children (n=68) attended mainstream elementary schools, and those who were identified as gifted were enrolled in special settings for gifted and talented children in the western part

of the Netherlands. Gifted children (n=45) were over-sampled and preliminary identification of giftedness took place on the basis of their enrolment in gifted education and qualitative judgements of parents and teachers regarding their giftedness. Schools participated on a voluntary basis, and written permission to participate was obtained from the children's parents and schools prior to participation. Six children dropped out in the course of the study, as they did not participate in each test session.

Design

The study utilized a 2 x 2 pre-test-post-test control group design with randomized blocks with Ability category (gifted versus average ability) and Condition (dynamic testing versus unguided practice) as variables (see Table 1). Blocking was based on the scores on the Raven Standard Progressive Matrices test (Raven, 1981), a visual inductive reasoning test, administered before the pre-test. All the children who had been identified as gifted had obtained Raven scores of at least the 90th percentile. Then, Raven scores were used, per Ability category, in order to ensure differences in initial reasoning ability were as small as possible across the dynamic testing and unguided practice conditions, to block children into the unguided practice (control static) testing condition or the dynamic testing condition. Children in the dynamic testing subgroups received training between pre-test 2 and post-test, whereas children in the unguided practice subgroups received an unrelated dot-to-dot control task of equal length between pre-test 2 and post-test.

Table 1. Overview over the design

		Dynamic testing (n=53)[1]		Unguided practice (n=60)	
		Gifted (n=22)	Average-ability (n=31)	Gifted (n=23)	Average-ability (n=37)
Prior to dynamic/static testing	Raven	x	x	x	x
	BRIEF	x	x	x	x
	BCST-64	x	x	x	x
Dynamic/static test	Pre-test 1	x	x	x	x
	Pre-test 2	x	x	x	x
	Dynamic training	Dynamic training	Dynamic training	Dot-to-dots control task	Dots-to-dots control task
	Post-test	x	x	x	x

[1] This study employed the same participants as in the study described in Chapter 3

The design included pre-test sessions 1 and 2 in order to enable comparisons between static and dynamic progression. During the pre-test sessions and the post-test, all children were only provided with short, general instructions and were not given any feedback. Administration of the instruments, including the training session, took approximately 20-30 minutes per session.

Materials

Raven. All participants were administered the Raven Standard Progressive Matrices Test (Raven, 1981) as a measure of their intellectual ability and a blocking instrument. The Raven test is a non-verbal intelligence test that measures fluid intelligence by means of multiple choice figural analogies. The Raven test results were shown to have a high level of internal consistency in several studies as shown by split-half-coefficients of r=.91 (Raven, 1981).

Berg Card Sorting Test-64 (BCST-64). The Berg Card Sorting Test-64 (Piper, Li, Eiwaz, Kobel, Benice et al., 2011), the shortened version of the BCST, was used to measure cognitive flexibility. The BCST is an open-source computerized version of the Wisconsin Card Sorting Test (WCST; Grant & Berg, 1948). Evaluation of the BCST-64 has shown a very strong relationship with the full version of the BCST (Fox, Mueller, Gray, Raber, & Piper, 2013) and the WCST (Piper et al., 2011), and is therefore considered an appropriate alternative to the WCST. The number of perseverative errors made during the administration of the BCST-64 were used as a measure of the participants' cognitive flexibility. Higher perseverative errors correspond with lower cognitive flexibility.

BRIEF. The teacher questionnaire of the Dutch version of the Behavior Rating Inventory of Executive Functions (BRIEF; Smidts & Huizinga, 2009) was utilized to obtain an approximation of the teachers' evaluation of children's metacognition. Scores on the BRIEF Metacognition Index were used to obtain the teacher's evaluation of each child's metacognition. Higher scores of the BRIEF are associated with more deviations from the norm, or impairment of executive functions. The Metacognition Index was found to have a high level of internal consistency (Cronbach's α=.95, Smidts & Huizinga, 2009).

Dynamic version of geometric analogies.

Pre-tests and post-test. The dynamic test used in this study was comprised of geometric visuo-spatial analogies of varying difficulty of the type A:B::C:D (see Figure 1 for an example of a difficult analogy item), Both the pre-tests, and the post-test consisted of 20 items of various difficulty, part of a test battery originally created by Hosenfeld, Van den Boom, and Resing (1997), and adapted by Tunteler et al. (2008). Six basic geometrical shapes were used in the construction of the analogies: squares, triangles, hexagons, pentagons, circles, and ovals. Each analogy was constructed by means of five possible transformations: changing

position, adding or subtracting an element, changing size, halving, and doubling. The test was administered as an open-ended paper-and-pencil test, and children had to draw their answers.

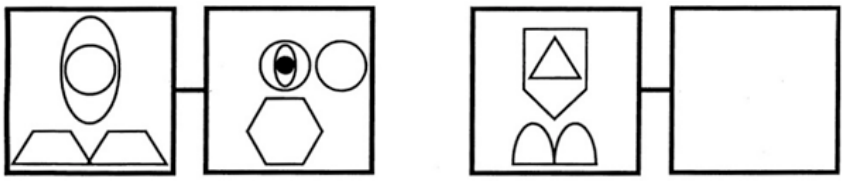


Figure 1. Example of a difficult analogy item.

The pre-tests and post-test, parallel sessions with different, but equivalent analogy items, were comprised of 20 trials with varying difficulty. The test sessions were equivalent in terms of the numbers of different elements, and transformations used for each analogy item, as well as the order in which the items were presented in relation to their difficulty level. The children received minimal instructions only in the two pre-tests and the post-test, as they were told that they had to solve puzzles with different shapes. The test leader then asked the child which shapes had to be drawn in the fourth box to solve the puzzle.

Training. The training session consisted of 10 geometric analogies that were not used in either the pre-tests or the post-test. The training session was based on graduated prompts techniques (Campione & Brown, 1987; Resing, 2000; Resing & Elliott, 2011). The prompts were administered following a standardized protocol, and were provided hierarchically, from two very general metacognitive prompts to two concrete cognitive prompts tailored to each specific item (see Appendix). Prompts were given if a child could not solve the analogy independently. After each prompt, children were asked to draw the solution of the analogy, and check their answer. If, after the fourth prompt, a child had not solved the analogy correctly, the test leader modelled the correct answer for the child. After the four prompts had been provided, and/or the test leader had shown the correct answer, the children were asked to explain why they thought their answer was correct. Then, the test leader provided a correct self-explanation. A schematic overview of the training procedure is included in the Appendix.

General procedure

The children were tested once a week over a period of five consecutive weeks. All tests and questionnaires part of the present study were administered following standard, protocolled instruction. At the beginning of the pre-tests, training session, and post-test, the children were provided with the six geometrical shapes used in the analogies, and in cooperation with the test leader named each shape, after which the test leader asked the child to draw the shapes below the printed shapes, staying as close to the original as possible.

Scoring

Analogy items were scored on the basis of children's drawings, in combination with their verbal explanations. Some of the children experienced difficulties drawing the geometrical shapes. This did not, however, cause any problems in scoring, as each child had copied the shapes used in the analogies on the cover sheet, so in the vast majority of cases the test leader knew which shapes the child had attempted the draw. In the few cases that it was, on first instance, unclear to the test leader which shape(s) the child had drawn, the child would be asked to point out on the cover sheet which shapes were intended.

For each item, the number of transformations that the child had applied correctly in solving the analogy was scored. Each analogy item was constructed by means of 1, 2, 3, 4, or 6 transformations that the child had to apply correctly in order to accurately solve the item, adding up to a total of 59 transformations per test session. The total number of transformations applied correctly in solving the analogies was taken as the outcome variable for each test session (Resing, Bakker, Pronk, & Elliott, 2016).

In order to estimate coding reliability, the pre-test 1 data were scored by both the first author and a student assisting in data collection. An inter-rater reliability analysis was performed using Cohen's κ to determine the level of agreement between the two raters. The inter-rater agreement for the pre-test 1 correct transformations was found to be very good, as determined by $\kappa=.83$, $p<.0001$.

Analyses

Multilevel modeling was used to analyze the current data. Multilevel modeling capitalizes on the hierarchical structure of the data, allowing us to study relations among variables at different levels and across levels. We can simultaneously answer level-1 questions about within-person change, and level-2 questions about how these changes vary across children (Singer & Willett, 2003). In the current study, level 1 represented the repeated measurements of the number of correct transformations within children, and level 2 represented the variability between children. We followed a predetermined model building structure as proposed by Singer and Willett (2003); starting with two simple, unconditional models and including our time-variant and time-invariant predictors in the successive models. The predictors were: condition, Ability category, cognitive flexibility and metacognition. Two time-invariant predictors, metacognition and cognitive flexibility, were mean centered to improve interpretation (Singer & Willett, 2003).

R (R Development Core Team, 2014) was used to fit the models. The fit of all models was compared using the likelihood ratio test (LRT) and two fit indices:

Akaike's Information Criterion (AIC) and the Schwarz's Bayesian Information Criterion (BIC). The likelihood ratio test follows a χ^2-distribution where the degrees of freedom are equal to the difference in the number of estimated parameters between the models. The LRT compares the "log likelihood" of two models and tests whether they differ significantly. The AIC and BIC are ad hoc criteria that are also based on the log likelihood statistic. The AIC and BIC statistics can be compared for all pairs of models, whether the models are nested within one another or not (Singer & Willett, 2003). These indices use a penalty function based on the number of parameters so that the more parsimonious model is favoured. A lower AIC and BIC value indicates a better fit of the model (Singer & Willett, 2003). All the discussed models were fitted using the Full Maximum Likelihood estimation. Most of the models differed in their fixed parts, and therefore deviance based on FML was needed to be able to compare the successive models (Singer & Willett, 2003).

4.3. Results

Before analysing the data for our research questions, one-way analyses of variance were conducted separately for each Ability category to evaluate possible differences between children in the two experimental conditions. The total Raven scores, pre-test 1 number of correct transformations, and age in months were used as dependent variables, and Condition with two levels (dynamic testing versus unguided practice) as independent variable. The findings for the gifted children revealed no significant differences in Raven scores (p=.53), pre-test 1 correct transformations (p = .40), nor in age (p=.52) between the dynamic testing and unguided practice conditions. Similarly, for the average-ability children no significant differences were found in Raven scores (p=.61), pre-test 1 correct transformations (p = .85), nor in age (p=.98) between the children in the two experimental conditions. We also examined possible differences between the gifted and average-ability children. The gifted children outperformed their peers on both the Raven scores, and the pre-test 1 correct transformations (for both measures, p<.001), but no significant differences were found in age (p=.31). Descriptive statistics of all measures used in the current study, per condition and Ability category are provided in Table 2.

Table 2. Mean scores and standard deviations of Raven scores, pre-test 1, pre-test 2, post-test correct transformations, cognitive flexibility and metacognition per condition and ability group

		Gifted		Average-ability	
		Dynamic testing	Unguided practice	Dynamic testing	Unguided practice
	N	22	23	31	37
Raven	M	43.82	44.57	34.55	33.78
	SD	4.22	3.78	5.53	6.47
Pre-test 1	M	39.14	41.96	29.16	28.43
	SD	15.13	9.26	13.56	15.77
Pre-test 2	M	46.86	53.74	43.52	41.03
	SD	17.62	4.05	13.40	18.27
Post-test	M	54.59	53.91	52.77	41.68
	SD	9.63	5.97	7.14	18.14
Cognitive flexibility	M	11.36	12.87	9.81	13.84
	SD	5.14	7.43	5.53	7.79
Metacognition	M	59.91	61.61	59.47	60.30
	SD	15.68	20.28	17.21	15.42

We conducted growth curve analyses (MLA) to model growth in the number of correct transformations. Table 3 presents the parameters and fit indices of the models. We first fitted the unconditional means model (intercept-only model) to acquire the random effects. The unconditional means model (Model 1) revealed a significant intercept effect (p<.001). We examined the intra-class correlation coefficient (ICC) as a measure of dependence; it describes the proportion of outcome variance that lies between persons in the population (i.e. the cluster structure of the data). As indicated by the intra-class correlation coefficient (ICC), of the total variation in the number of correct transformations, 54.38% could be attributable to differences between children. This finding revealed that the observations were not independent, and indicated that there was systematic variation in the outcome measure (transformations) worth exploring, both for the within-level and between-level variance, reinforcing the choice of multilevel modelling.

In Model 2 (the unconditional growth model), we included our time predictor into the level-1 sub-model in order to explain the remaining within-child variance (117.8). The estimated rate of change in the number of correct transformations for an average participant was 8.13 (p<.001); children generally improved in the number of correctly applied transformations. A negative covariance (-0.56)

was found between the slope and intercept. This indicated that children using fewer correct transformations at pre-test 1 increased their number of correct transformations slightly faster across test sessions than children with a higher number of correct transformations at pre-test 1. Variance components revealed remaining variance in the number of correct transformations both between, and within, children. Extending the model by adding other predictors could possibly reduce this variation.

Model 3 included Condition as an explanatory variable for the number of correct transformations. Result of the likelihood ratio test (LRT) showed that model fit improved ($X^2(1)=5.46$, $p=.02$). Children of the unguided practice group had, on average, an estimated rate of change of 7.31. Therefore, these children generally increased their number of correct transformations across test sessions. A positive fixed effect for Condition (training versus unguided practice) of 3.51 revealed that the dynamic training session influenced the performance of the children. In accordance with our expectation, those who received a dynamic training session improved more in the number of correct transformations from pre-test 2 to post-test than the children in the unguided practice condition.

In Model 4 we included Ability category, gifted versus average-ability, as a predictor for initial status. Model 4 provided a better fit to the data compared to Model 3 ($X^2(1)=10.82$, $p=.001$). Children's Ability category was found to be related to the number of correct transformations at pre-test 1 as shown by a significant main effect of Ability category (8.23). Specifically, children with higher intellectual ability scored, on average, higher on pre-test 1 than average-ability peers. Model 5 showed that Ability category was also a significant predictor for children's rate of change, as indicated by a significant interaction of Ability category and Time. Model fit improved ($X^2(1)=4.96$, $p=.03$). The estimate (-2.21) revealed that average-ability children improved more in the number of correct transformations over time than gifted children.

In Model 6 we examined whether the dynamic training session had different benefits for gifted and average-ability children. We included the interaction effect of Ability category and Condition, which did not improve model fit ($X^2(1)=1.75$, $p=.19$). No significant difference was found in dynamic training benefits for gifted and average-ability children, as revealed by the non-significant interaction effect (-3.85), indicating that gifted children did not show more progression in the number of correct transformations after training than their average-ability peers.

Model 7 showed no significant main effect of Cognitive flexibility; model fit did not improve ($X^2(1)=0.53$, $p=.47$). The non-significant interaction effect of Cognitive flexibility x Time in Model 8 ($X^2(2)=0.59$, $p=.75$) indicated that we could

not support our expectation that children with higher levels of cognitive flexibility would show more progression in the number of correct transformations than their age-mates with lower levels of cognitive flexibility. Children with higher levels of cognitive flexibility did also not benefit more from the dynamic training session than children with lower levels of cognitive flexibility as shown in Model 9 ($X^2(2)=2.84$, $p=.24$). Furthermore, results of Model 10 showed that the progression paths of gifted children that had higher levels of cognitive flexibility were not steeper than those of their average-ability peers ($X^2(2)=2.47$, $p=.29$). The time-invariant predictor Cognitive flexibility was not included in the remaining models.

Model 11 included the main effect of Metacognition. A non-significant effect was found, however, model fit did improve after inclusion of the predictor ($X^2(1)=22.80$, $p<.001$). Results of Model 12 showed that children with higher scores on the Metacognition Index showed equivalent progression in the number of correct transformations across test sessions than their peers with lower scores on the Metacognition Index ($X^2(1)=2.97$, $p=.08$). In Model 13, we included the interaction effect of Metacognition and Condition, which led to an improvement in model fit ($X^2(1)=4.40$, $p=.04$). The estimate (0.149) showed that children with higher scores on the Metacognition Index benefited more from training than peers with lower scores. We included the three-way interaction between Condition, Ability category and Metacognition in Model 14. Results showed that the progression paths of gifted children that had higher levels of metacognition were not steeper than those of their average-ability peers ($X^2(1)=0.20$ $p=.66$).

In conclusion, Model 13 was shown to be the model that best fitted the data based on the LRT, and the AIC and BIC statistics. The dynamic sessions led to an improvement in the number of correct transformations the children used. No differences in dynamic training benefits for gifted and average-ability children were found. The average-ability children in the unguided practice condition did, however, show more improvement across test sessions than the gifted children in the unguided practice session. Cognitive flexibility did not influence children's progression over time and the improvement in the number of transformations after receiving the dynamic training. The progression paths did also not differ for gifted children with higher levels of cognitive flexibility and their average-ability peers. Metacognition did not influence progression in the number of correct transformations. Children with higher scores on the Metacognition Index, indicating lower levels of metacognition, showed more improvement in the number of correct transformations after the dynamic training than their peers with lower levels of metacognition. Lastly, the progression paths did not differ between gifted children who had higher levels of metacognition and their average-ability peers.

Table 3. Results of the fitted multilevel models for the number of correct transformations

Model	Estimate*(SE)*	Deviance	AIC	BIC
1. Intercept only	42.89(1.26)**	2750.6	2756.6	2768.1
2. Time	8.13(0.51)**	2557.8	2569.8	2592.7
3. Condition	3.51(1.40)*	2552.3	2566.3	2593.1
4. Ability category	8.23(2.39)**	2541.5	2557.5	2588.1
5. Ability category x Time	-2.21(0.98)*	2536.5	2554.5	2589.0
6. Ability category x Condition	-3.85(2.82)	2534.8	2554.8	2593.1
7. Cognitive flexibility	-0.13(0.17)	2536.0	2556.0	2594.3
8. Cognitive flexibility x Time	0.02(0.07)	2536.0	2558.0	2600.0
9. Cognitive flexibility x Condition	0.34(0.21)	2533.7	2555.7	2597.8
10. Cognitive flexibility x Condition x Ability category	0.49(0.35)	2534.1	2556.1	2598.2
11. Metacognition	-0.03(0.07)	2513.7	2533.7	2571.9
12. Metacognition x Time	0.05(0.03)	2510.8	2532.8	2574.8
13. Metacognition x Condition	0.15(0.07)*	2509.3	2531.3	2573.3
14. Metacognition x Condition x Ability category	-0.06(0.14)	2509.1	2533.1	2578.9

Note. Significance: ** $p < .001$, * $p < .05$. The deviance, AIC, and BIC statistics were examined for the relative goodness-of-fit of the successive models.

In order to examine our final research question regarding potential differences in the instructional needs of gifted and average-ability children, we conducted a one-way ANOVA with two within-subjects factors (metacognitive and cognitive prompts) and one between-subjects (Ability category) factor with the number of prompts in each category as dependent variables. No significant differences were found in the number of metacognitive, $F(1,51)=2.27$, $p=.14$, or cognitive prompts, $F(1,51)=.17$, $p=.69$ across ability categories (see Table 4).

Table 4. Mean scores and standard deviations of the number of metacognitive and cognitive prompts received during training per Ability category

	Metacognitive prompts		Cognitive prompts	
	M	SD	M	SD
Gifted	11.91	2.14	2.41	4.47
Average-ability	12.87	2.39	2.90	4.29

4.4. Discussion

The current study explored the potential differential benefits of dynamic versus static testing of gifted and average-ability children, and focused on two aspects of executive functioning, cognitive flexibility and metacognition. First of all, our results showed that children who had unguided practice experience only, and children who were dynamically tested showed progression in the number of correct analogical transformations. When children were tested dynamically, however, their progression paths were shown to be more advanced, which supports previous findings (Resing, 2000; Stevenson et al., 2013, 2014). In this sense, our findings build upon earlier studies in which it was posited that dynamic testing of children reveals a more complete picture of their cognitive potential than static testing only (e.g., Elliott, 2003; Sternberg & Grigorenko, 2002).

Moreover, our findings indicated, as expected, that gifted children start at a higher ability point, and keep this advantage during following sessions. When looking into potential differences between gifted and average-ability children in relation to the nature of progression, in contrast to our expectations, it was found that, in general, the average-ability children showed more progression than their gifted peers. We cannot, however, discount that the gifted children in the current study might have experienced a ceiling effect in testing, which could have influenced the research results. If these children had indeed experienced a ceiling effect, we would then have expected them to show a differential need for instructions, which could not be supported by our data. Moreover, neither the original authors of the items used in the current study (Hosenfeld et al., 1997), nor others who have used these items (e.g., Tunteler et al., 2008) for children of the same age report on a ceiling effect. It must be mentioned, nevertheless, that it is not known whether any high-ability children participated in these studies. Therefore, this explanation requires further research.

Looking more closely into training benefits, it was revealed that the gifted and average-achieving children showed similar rather than different progression lines after training, whereas previous studies into dynamic testing of gifted children found that these groups of children differed significantly in their performance and progression (e.g., Calero et al., 2011; Kanevsky, 2000; Kanevsky & Geake, 2004). Although we cannot completely discount a potential ceiling effect, as described above, in the light of the fact that all groups of children progressed after training, our findings, ultimately, seem to suggest that dynamic testing might be better suited to reveal children's cognitive potential of all groups of children (Elliott et al., 2010), including those with above-average cognitive abilities.

We also examined the role that cognitive flexibility and metacognition play in progression in accuracy of analogical reasoning, and training benefits.

It could not be established that cognitive flexibility plays a role in progression of analogical reasoning or training benefits. A number of reasons can be identified for the unexpected results regarding cognitive flexibility. First of all, research into executive functioning amongst children is challenging. One important reason is the type of instruments used to measure executive functioning. It has been noted that performance-based tasks, such as the BCST-64 used in the current study, rarely measure one executive function only (e.g., Miyake, Friedman, Emerson, Witzki, Howerter et al., 2000). By definition, executive functions regulate various cognitive processes, including for instance visuospatial processing. Performance-based tasks measure these other processes as well, making measuring just one executive function, in isolation, difficult (Viterbori et al., 2015). The developmental nature of executive functions in childhood should also be taken into consideration (e.g., Diamond, 2013; Kuhn, 2000). Moreover, it should be noted that the cognitive flexibility task used in the current study is a single measurement, static test, whereas learning potential measures are dynamic. Therefore, future studies could research this relationship further by utilizing a dynamic cognitive flexibility task, such as the dynamic Wisconsin Card Sorting Task (e.g., Boosman, Visser-Meily, Ownsworth, Winkens, & Van Heugten, 2014). These authors found that the dynamic executive functioning indices were significantly associated with cognitive functions, whereas the static indices were not.

It was, nonetheless, found that metacognition had an effect on the training benefits, but not on the progression from pre-test to post-test. Although it was expected that children with higher levels of metacognition would benefit more from training, we found the reverse. Children who, according to their teachers, had lower levels of metacognition benefitted more from training than their peers with higher levels of metacognition. This finding, once more, shows how dynamic testing can reduce test bias, and, in that way, lead to profound insights into how children learn (e.g., Elliott et al., 2010). Furthermore, the findings provide a first indication that a graduated prompts training procedure can, to a certain extent, compensate for lower levels of metacognition. This notion is particularly relevant considering Sternberg's (1998) assertion that metacognition is an important ability in the development of expertise.

Although it seems plausible that the graduated prompts technique used in the current study also helps improve metacognition, this tentative hypothesis cannot be asserted in the current study, and should be investigated using several measurements of metacognition. It must be noted that, although studies suggest that rating scales can be used successfully in order to obtain an approximation of children's executive functioning (Toplak, West, & Stanovich, 2013), using teacher ratings is a very indirect method of measuring metacognition. Future studies

should therefore also focus on development and implementation of instruments that directly measure or predict executive functioning amongst young children.

Finally, we looked more closely into children's instructional needs during dynamic training. Contrary to what we expected based on previous literature (e.g., Calero et al., 2007; Kanevsky & Geake, 2004), we found no differences in the instructional needs of the gifted versus average-ability groups of children. Individual differences between children's need for instructions, both within and across ability categories, were, however, found, which is in line with previous studies (e.g. Resing, 2013; Sternberg & Grigorenko, 2002). Our study, moreover, suggests that children, regardless of whether they have high or average levels of cognitive abilities, can have a similar need for instructions in order to progress in learning. Of course, follow-up studies are required in order to investigate whether these findings are domain-specific or general.

In addition to the limitations mentioned above regarding measuring executive functioning and a potential ceiling effect, the current study encountered some other limitations. First of all, it is important to mention that we only used the Raven Standard Progressive Matrices as a measure of intellectual ability. Although the Raven test is known as a robust measure of intellectual ability (e.g., Jensen, 1998), we did not include other factors deemed important for cognitive and intellectual functioning, such as task commitment or creativity (e.g., Renzulli, 2005; Renzulli, & D'Souza, 2014). Moreover, we only investigated correct analogical transformations, while other factors have also been shown to be important in progression in analogical reasoning. Investigating strategy use, in particular, could lead to interesting findings considering the assumed relationship between strategy use and aspects of executive and intellectual functioning (e.g., Shore, 2000).

The results of the current study yield some important implications for educational professionals. In the context of the current study, it seems advisable to administer a dynamic rather than a static test when children's intellectual abilities are questioned, especially for children with lower levels of metacognition. Not only do our results underline the notion posited in a myriad of earlier work (e.g., Elliott et al., 2010; Resing, 2000, 2013; Sternberg & Grigorenko, 2002) that static testing does not always show a full picture of children's cognitive potential, our findings also indicate that children with different levels of intellectual ability, including those who have the potential to excel, can profit from dynamic testing, and, in particular, that children with lower levels of metacognition benefit more from training than their peers with higher levels of metacognition. Ultimately, the latter finding suggests that dynamic testing, in particular, may result in a more accurate view of the cognitive abilities of children with lower levels of metacognition.

Opponents of dynamic testing often argue that testing dynamically is more labour-intensive, and, thus, more expensive than testing statically. Nevertheless, as the children in the two ability categories showed progression after unguided repeated practice, and, more importantly, steeper progression lines after dynamic training, these findings suggest that gifted children also learn within the zone of proximal development (e.g., Calero et al., 2011). It seems that taking extra time to test these children more than once and administering a dynamic training session, helps them in unveiling their cognitive abilities, and, thus, is worth the extra investment.

This notion becomes even more salient when taking into account that dynamic testing of children also provides insight into their instructional needs (e.g., Bosma & Resing, 2012). The results of the current study indicate that children of different levels of intellectual ability, including those with the potential to excel, can have a similar need for instructions, and can profit from similar help. Furthermore, our findings remind us that, when teaching high-ability children, these children do not, by definition, need less instruction or feedback than average-ability children, in order to show progression in learning. Just like any other children, some of these children can also profit from extra feedback or help so they can unveil their true cognitive potential. Finally, and most importantly, the results of the present study indicate that children, even those who have already achieved excellent results, can show learning progression when they are provided with the right instructions.

Appendix. Schematic overview of the graduated prompts training protocol

STEP	INSTRUCTION	INCORRECT ANSWER?	CORRECT ANSWER?
1	This is another puzzle with four boxes. Do you remember what we are going to do? *(have child provide an answer)* We are going to solve the puzzle by filling the empty box with the correct figures. Just draw the answer that you think is correct in the empty box *(have child draw the answer)*. Check whether you drew the correct answer *(have child check and correct answer if necessary)*	The picture you drew is great, but it is not entirely correct yet. I will help you, but try to find the correct answer with as little help from me as possible. We will start again after each try.	To step 5: Well done, that is the correct answer! Can you tell me why this this the correct answer? *[Test leader models correct self-explanation, as per the protocol, tailored to each item]*
2	How do we start? *(have child provide an answer)* First, have a good look at the figures in these three boxes *(point at A, B, C)* Do you now know the correct answer? Just draw the answer that you think is correct in the empty box *(have child draw the answer)* Check whether you drew the correct answer *(have child check and correct answer if necessary)*	Great picture! It is not entirely correct. I will help you some more.	
3	Have a good look at these boxes *[point at A and B]* What do you see? *[Have child provide an answer]* We see that A and B belong together. Do you know why? *[have child provide an answer]* *[Then explain the transformations from A → B according to protocol, tailored per item]* Do you now know the correct answer? Just draw the answer that you think is correct in the empty box *(have child draw the answer)* Check whether you drew the correct answer *(have child check and correct answer if necessary)*	You drew another beautiful picture. It is almost correct, so I will help you a little bit more.	
4	Now have a good look at this box *[point at C]* and this box *[point at A]* What do you see? *[Have child provide an answer]* We see that A and C look alike, but that they changed a little bit. Can you tell me why? *[Have child provide an answer]* *[Then explain the similarities between A and C, B according to protocol, tailored per item]* Do you now know the correct answer? Just draw the answer that you think is correct in the empty box (have child draw the answer) Check whether you drew the correct answer *(have child check and correct answer if necessary)*	What a beautiful picture. You can draw very well. It is not entirely correct; I will show you the correct answer *[test leader draws correct answer]* Can you tell me why this this the correct answer? *[Test leader models correct self-explanation, as per the protocol, tailored to each item]*	