Computerised dynamic testing: An assessment approach that tailors to children's instructional needs

Touw, K.W.J.

Chapter 3

Progression and individual differences in children's series completion after dynamic testing

Kirsten W. J. Touw
Bart Vogelaar
Floor Thissen
Sanne Rovers
Wilma C. M. Resing

Touw, K. W. J., Vogelaar, B., Thissen, F., Rovers, S. & Resing, W. C. M. (2020). Progression and individual differences in children's series completion after dynamic testing. *British Journal of Educational Psychology.* https://doi.org/10.1111/bjep.12272

Abstract

Background. The need to focus more on children's abilities to change requires new assessment technologies in education. Process-oriented assessment can be useful in this regard. Dynamic testing has the potential to provide in-depth information about children's learning processes and cognitive abilities.

Aim. This study implemented a process-oriented dynamic testing procedure to obtain information regarding children's changes in series-completion skills in a computerised test setting. We studied whether children who received a graduated prompts training would show more progression in series-completion than children who received no training, and whether trained children would use more advanced explanations of their solutions than their untrained peers.

Sample. Participants were 164 second-grade children with a mean age of 7;11 years. Children were split into an unguided practice or a dynamic testing condition.

Methods. The study employed a pre-test-training-post-test design. Half of the children were trained in series-completion, and the other half did not receive any feedback on their problem solving. Using item response theory analysis, we inspected the progression paths of the children in the two conditions.

Results and conclusions. Children who received training showed more progression in their series-completion skills than the children who received no training. In addition, the trained children explained their solutions in a more advanced manner, when compared with the non-trained control group. This information is valuable for educational practice as it provides a better understanding of how learning occurs and which factors contribute to cognitive changes.

3.1 Introduction

One of the focal points in education is helping students make the most of their learning. Teachers are repeatedly asked to improve students' learning and cater to their individual educational needs. As part of the discussion around enhancing learning opportunities, Gotwals (2018) suggested that incorporating formative assessments within the classroom is the way forward. Formative assessment tools provide feedback to teachers to help students learn more effectively, as a consequence improving students' academic achievements (Dixson & Worrell, 2016). Despite the widely recognized need for schools to focus on personalisation and learning how to learn, education is still dominated by assessment and testing practices that focus on the summative assessment of learning outcomes, rather than on formative assessment practices that support and strengthen students as learners (Bennett, 2011; Crick, 2007). The need to focus more on students' abilities to change requires the development of new assessment technologies. Process-oriented assessment techniques, such as dynamic testing, can be useful in this regard.

A dynamic testing approach has the potential to provide in-depth information about children's learning processes and cognitive abilities (Elliott, Resing, & Beckman, 2018). This information can be used to develop effective educational practices (Elliott, Grigorenko, & Resing, 2010; Jeltova et al., 2007). Our study aimed to address the need for new assessment technologies that can be used to obtain more insight into children's learning processes. We have newly constructed a computerized series-completion test in a dynamic testing setting, to better be able to assess children's progression in solving a domain-general inductive reasoning task.

Computerized dynamic testing

Recently, the benefits of adding electronic technology to a dynamic testing design has been examined by several researchers (e.g., Passig, Tzuriel, & Eshel-Kadmi, 2016; Poehner & Lantolf, 2013; Resing & Elliott, 2011; Stevenson, Touw, & Resing, 2011). Incorporating electronic displays is believed to contribute to the development of children's cognitive skills (e.g., Clements & Samara, 2002). The additional value of computerized testing can be attributed to the flexibility with which problems can be solved, which can promote more adaptive prompting during training. Research has shown that children benefit from computer-assisted learning (Tamim, Bernard, Borokhovski, Abrami, & Schmid, 2011), and computerized dynamic testing has shown positive results in relation to children's accuracy on cognitive tasks (e.g., Passig et al., 2016; Poehner & Lantolf, 2013; Resing & Elliott, 2011; Resing, Steijn, Xenidou-Dervou, Stevenson, & Elliott, 2011; Stevenson et al., 2011; Tzuriel, & Shamir, 2002). In the current study, we developed a computerized, tablet-based dynamic test of inductive reasoning, which enabled us to examine the following two aims. Firstly, using a dynamic test allowed us to investigate children's ability to learn. Secondly, we aimed to develop a digital test that could potentially

be used in education as a first step for developing a more effective and integrated learning environment. Moreover, computerized dynamic testing not only allows for the investigation of emerging individual differences during the process of solving cognitive tasks, but also provides information about factors that influence performance change (Elliott et al., 2018).

Dynamic testing: Measuring change in children's accuracy

The dynamic testing approach draws, among others, upon Vygotsky's theory of the zone of proximal development (ZPD) (Vygotsky, 1978), which has been influential in education (Elliott et al., 2018). Dynamic tests examine the changes that occur during an assessment (Tzuriel, 2011) by incorporating feedback and training into the testing phases, providing information about the individual's ZPD. The design of traditional static assessment methods does not allow for discriminating between what a child can achieve with and without help (Elliott et al., 2018). By tapping into underlying potential rather than the current unaided abilities, however, dynamic testing does more than merely examine the present cognitive abilities of children (Elliott et al., 2010; Grigorenko, 2009; Haywood & Lidz, 2007). By focusing on developing abilities and providing instruction or help as part of the testing procedure, these tests, potentially, provide insight into children's cognitive potential, or potential for learning (Hill, 2015; Tiekstra, Minnaert, & Hessels, 2016).

A training procedure utilized in dynamic testing involves the provision of graduated prompts (e.g., Campione & Brown, 1987; Ferrara, Brown, & Campione, 1986; Resing, 1997; Resing & Elliott, 2011). This standardized method, based on the concept of differing degrees of help, comprises provision of prompts in a gradual, hierarchic fashion when independent problem-solving does not lead to an accurate solution. As the provision of prompts is determined by the child's needs, this training approach is believed to provide more information about a child's problem-solving process than standardized, conventional testing (Resing, 2013).

For decades, however, researchers have debated about the best way of measuring change in dynamic testing (e.g., Cronbach & Furby, 1970; Harris, 1963). In particular, the reliability of gain scores in a pre-test-training-post-test design have been criticized because of the possibility of ceiling effects and regression to the mean; whereby, a progression in scores of, for example, 4 points from 1 to 5 items can have a different meaning than a progression from 13 to 18 points for a test of 20 items (e.g., Guthke & Wiedl, 1996). To overcome the limitations of classical test theory, Item response theory (IRT) was utilized in this study. IRT models enable estimating the probability of solving an item correctly, based on the child's ability and the item difficulties (e.g., Embretson, 1987, 1991; Embretson & Prenovost, 2000; Embertson & Reise, 2000). In this way, these models provide a more favourable reliability of gain scores and their interpretation within a dynamic testing context (Stevenson, Heiser, & Resing, 2016; Stevenson, Hickendorff, Resing, Heiser, & De Boeck, 2013). Hessels and Bosson (2003) and De Beer (2005) also used Rasch scaling in dynamic testing with the HART, and the Computer

Adaptive Test of Learning Potential, respectively. In the current study, we therefore used IRT-based gain scores to measure children's performance changes at the group level.

Children's verbal explanations of their series-completion task solving

Another important component of children's performance changes is their use of solving strategies (Siegler & Svetina, 2002). By examining the changes in children's ways of solving the tasks throughout the test sessions, it would be possible to analyse in-depth the learning processes that may have occurred (Siegler, 2007; Siegler & Svetina, 2006). One way of looking into these solving strategies is to study children's verbal explanations, in which they explain how they solved a task (Farrington-Flint, Coyne, Stiller, & Heath, 2008; Pronk, 2014; Resing, Xenidou-Dervou, Steijn, & Elliott, 2012; Siegler & Stern, 1998). These verbal explanations provide information about children's strategies and problem-solving knowledge and seem to have good validity (Reed, Stevenson, Broens-Paffen, Kirschner, & Jolles, 2015; Taylor & Dionne, 2000). In relation to dynamic testing, Resing and colleagues (2012) and Resing, Bakker, Pronk and Elliott (2016), for example, found that children's verbal problem-solving strategies regarding a series-completion task progressed to a more advanced level of reasoning after dynamic training. These trained children became better at explaining the separate item attributes and how these changed in the series they had to solve, when compared with their non-trained peers.

In the current study, we investigated two aspects of children's performance changes: changes in accuracy in solving inductive reasoning tasks and changes in their verbal explanations.

Factors influencing individual differences in task solving

Substantial interindividual differences have been observed in the extent to which children show progression in task solving (Tunteler, Pronk, & Resing, 2008). Several studies in dynamic testing showed that children with a low initial ability profited more from training in inductive reasoning than children with a higher initial ability (e.g., Stevenson, Hickendorff, et al., 2013; Swanson & Lussier, 2001). Also, working memory has been hypothesized to contribute to children's performance during dynamic testing (e.g., Resing, Bakker, Pronk, & Elliott, 2017; Resing et al., 2011; Stevenson, Bergwerff, Heiser, & Resing, 2014). Earlier research on dynamic testing has reported that both verbal and visual-spatial working memory components play a role in solving visual-spatial analogies. This is particularly apparent when, as part of the assessment, children are asked to explain their problem-solving procedures (Resing, Bakker, et al., 2017; Stevenson, Heiser, et al., 2013; Tunteler et al., 2008).

Aims of the current study

This study's main aim was to examine children's ability to progress in solving geometric series-completion items, after they were provided with feedback in task solving, provided by a tablet. We thereby focused on children's potential improvement in accuracy of task solving and their verbal explanations. Rasch scaling based on Embretson's IRT modelling was utilized to study children's progression from pre-test to post-test in series-completion accuracy, that is gain scores. On the basis

of earlier findings about the effect of dynamic testing on children's accuracy, it was expected that trained children would improve their reasoning accuracy, as measured by their gains, more than the control-group children (e.g., Resing, Touw, Veerbeek, & Elliott, 2017). We also expected that dynamically trained children would employ more sophisticated verbal explanations at the post-test in comparison with the pre-test explanations than the untrained control group (Resing et al., 2016).

Moreover, we studied some factors that would potentially influence individual differences in solving series-completion task-items, by inspecting interindividual differences in performance changes between the pre-test and post-test stages. Previous research on inductive reasoning has focused on working memory (e.g., Resing, Bakker, et al., 2017; Stevenson, Heiser, et al., 2013; Swanson, 2011) and initial ability (e.g., Stevenson, Hickendorff, et al., 2013). On the basis of these earlier study results, we explored whether these factors would influence dynamic test outcomes.

3.2 Method

Participants

The participants in this study were 164 second-grade children, 89 girls and 75 boys, ranging in age from 6 years and 7 months to 9 years and 3 months (*M* = 94.91 months, *SD* = 4.9 months). The children were recruited from 14 primary schools, located in midsize and large towns in the western part of the Netherlands. The children's primary language spoken at school was Dutch. First, a random selection of regular primary schools in the vicinity of the research institution was contacted by phone and sent an information letter. If they agreed to participate, headmasters signed an informed consent form. Then, parents were informed, and written parental consent for participation was obtained for all children. Distribution of children throughout the participating schools was based on parents' signed consent, with a mean of 12.61 children (*SD* = 6.97) per school. Initially, the study included 177 children. However, 13 children dropped out in the course of the study because they had been absent during one or more of the testing sessions. No further exclusion criteria were applied. The research project was approved by the ethics board of our university.

Design

The study employed a control-group design consisting of pre-test, training, and post-test segments (see Table 1). Each child took part in five individual weekly sessions, separated by approximately 7 days. We used randomized blocking to avoid differences in initial reasoning ability between the two conditions. Blocking was based on children's scores on the Raven's Standard Progressive Matrices test (Raven, Raven, & Court, 1998) and the schools the children attended. Per school, blocks of two children were randomly allocated to the training or the control condition. Children completed a static pre-test that measured their initial abilities, in which they solved a series-completion test without feedback on their performance. Children in the training condition then

received two consecutive dynamic training sessions, followed by a post-test. Children in the control group solved mazes and dot-to-dot completion tasks between pre- and post-test, so that the contact moments with the test leader and the time-on-testing would be as equal as possible between the two groups.

Table 1. Schematic overview of the design of the study[1]

Condition	*N*	Raven	Pre-test	Training 1	Training 2	Post-test
Training	80	Yes	Yes	Yes	Yes	Yes
Control	84	Yes	Yes	No/mazes	No/mazes	Yes

Materials

Raven's progressive matrices. This is a nonverbal test (Raven et al., 1998) that measures children's fluid intelligence, especially their inductive reasoning. Children were asked to complete 60 multiple-choice items by choosing the missing element of a figure. The Raven test has a reliability of α = .83 and a split-half coefficient of *r* = .91 (Raven, 1981).

Automated working memory assessment (AWMA): Listening Recall. The Listening Recall subtest of the AWMA (Alloway, 2007) was used to measure children's verbal working memory. In this subtest, a child had to listen to a certain number of sentences and indicate whether these are true or not true. Next, the child had to repeat the first words of the sentences in the correct order. The reported test-retest reliability is *r* = .88 (Alloway, 2007).

Automated working memory assessment: Spatial recall. Visual-spatial working memory was assessed by the Spatial Recall subtest of the AWMA (Alloway, 2007). Children were shown two figures and had to indicate whether the second figure was the same as or the reverse of the first figure. In addition, the second figure contained a red dot. After inspecting a certain number of figures, the children had to recall the positions of these dots in the correct order. Alloway (2007) reported a test-retest reliability of *r* = .79.

Computerized dynamic test of series-completion: Construction. A new computerized series-completion test, utilizing geometric series-completion items, was used to measure children's inductive reasoning ability. In this task, children were asked to complete sequential patterns. A series of six boxes

[1] The studies described in chapter 2 and 3 made use of a shared dataset concerning computerized dynamic testing, but focused on different research questions.

filled with geometric figures and one empty box was presented. The children were asked to determine which figure was needed to complete the series and verbalize why they thought their solutions were correct. Determining the correct solution required discovering the number of pattern transformations and the period of change (periodicity) (Resing, Tunteler, & Elliott, 2015; Simon & Kotovsky, 1963). Discovering periodicity involves noticing that patterns are repeated at predictable, regular intervals (Holzman, Pelllegrino, & Glazer, 1983). The task has been constructed with items having a large range of (theoretical) difficulty levels depending on the number of transformations and the period of change in the items. Five transformations were possible: changes in geometric shape (circle, triangle, or square), colour (orange, blue, pink, or yellow), size (large or small), quantity (one or two), and positioning in the box (top, middle, or bottom). See Figure 1 for an example-item of the series-completion test.

Pre-test task difficulty for the sample of children in the current study, the mean *p*-value and range, was .42 (range .00 to .95) and .43 (range .01 to .96), for the control and dynamic training groups respectively. For the post-test, the mean *p*-value was .44 (range .02 to .95) and .59 (range .01 to 1.00) for the control and dynamic training groups respectively. A higher *p*-value shows more children solved the item correctly.

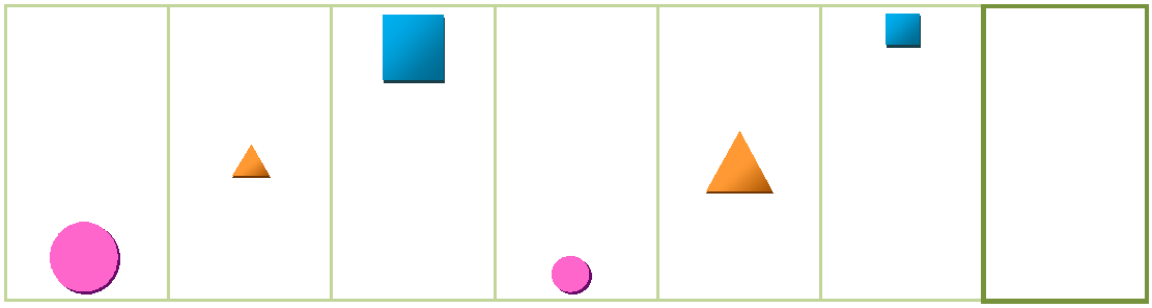

Figure 1. Geometric series-completion test. Item with four transformations: geometric shape (periodicity 3), colour (periodicity 3), size (periodicity 2), and position (periodicity 3)

Computerized dynamic test of series-completion: Pre-test and post-test. After two examples, 18 geometric series-completion task items were presented on a tablet in both the pre-test and post-test. The sessions comprised items equivalent in structure; the items had identical patterns of item difficulty but differed in the figures and colours that were used in the series. Before the start of the pre-test, the geometrical shapes used in this task were introduced to the children. Thereafter, the procedures of the pre-test and post-test were the same. Each session lasted approximately 30 min.

Internal consistency for the pre-test was α = .64. Post-test reliability for the control and the training conditions was α = .63 and α = .64, respectively. Test-retest reliability between the pre-test

and post-test scores for the children in the control group was found to be $r = .74$, $p < .001$. For the children in the training group, the test-retest reliability score was, as expected, lower: $r = .35$, $p = .002$.

Computerized dynamic test of series-completion: Training procedure. The two training sessions each consisted of six series-completion items that were comparable to those used in the pre-test and post-test. The order of the items presented during the training sessions ranged from difficult to easy. After a correct answer was provided during the training sessions, the children received positive feedback and were asked why they had chosen this answer. After an incorrect answer, graduated prompts (e.g. Campione & Brown, 1987; Ferrara et al., 1986; Resing, 1997; Resing & Elliott, 2011) were provided. The predetermined prompts ranged from general to specific instruction (see Figure 2). If a child could not solve the task independently, he or she was gradually prompted towards the correct solution, starting with general, metacognitive prompts. Subsequently, a more explicit, cognitive prompt that emphasized the specific transformations in the series was provided. If the child still could not accurately solve the task, direct guidance by scaffolding was provided.

Electronic device: Tablet

The task was presented on an Acer Aspire Switch 10 convertible tablet. This tablet operated on Windows and had a 10.1-inch touchscreen display with a resolution of 1,280 x 800 pixels. During the task, the tablet provided different kinds of output. On the tablet's display, an animated figure, named Lisa, appeared on the left side of the screen and gave the children verbal instructions. The children were asked to construct their answers by dragging and dropping geometric figure(s) (from a range of possibilities) into the empty seventh box. The possibilities (24 figures) were presented below the row of figures (see Figure 3). In addition, the tablet provided visual effects parallel to the verbal instructions in all four sessions to visually attract attention to the figures. The tablet briefly enlarged the geometric figures in the series, the outlines of the boxes, and the outline of the entire row. Furthermore, during the example and training items, the tablet provided auditory feedback. A high 'pling' sound was played whenever an answer was correct and a lower sound when the child's answer was incorrect. Appendix A presents a schematic and detailed overview of the computerized series-completion test presented on the tablet.

Scoring and analyses

The tablet automatically scored children's performance during the pre-test, training, and post-test by producing log files. For each of the 18 pre-test and post-test items, answers were scored as accurate (1) or inaccurate (0). To examine the effect of training on series completion performance, we used Embretson's (1991) multidimensional Rasch model for learning and change (MRMLC) to reliably estimate initial ability and change from pre-test to post-test (e.g., Embretson & Prenovost, 2000). Following Stevenson, Hickendorff, and colleagues (2013) we included condition as a covariate in our model to examine the effect of condition and reliably estimate change scores for each experimental

condition. Initial analyses were performed using the ltm package for R (Rizopoulos, 2006); MRMLC estimates were computed with the lme-4 package (Bates & Maechler, 2010).

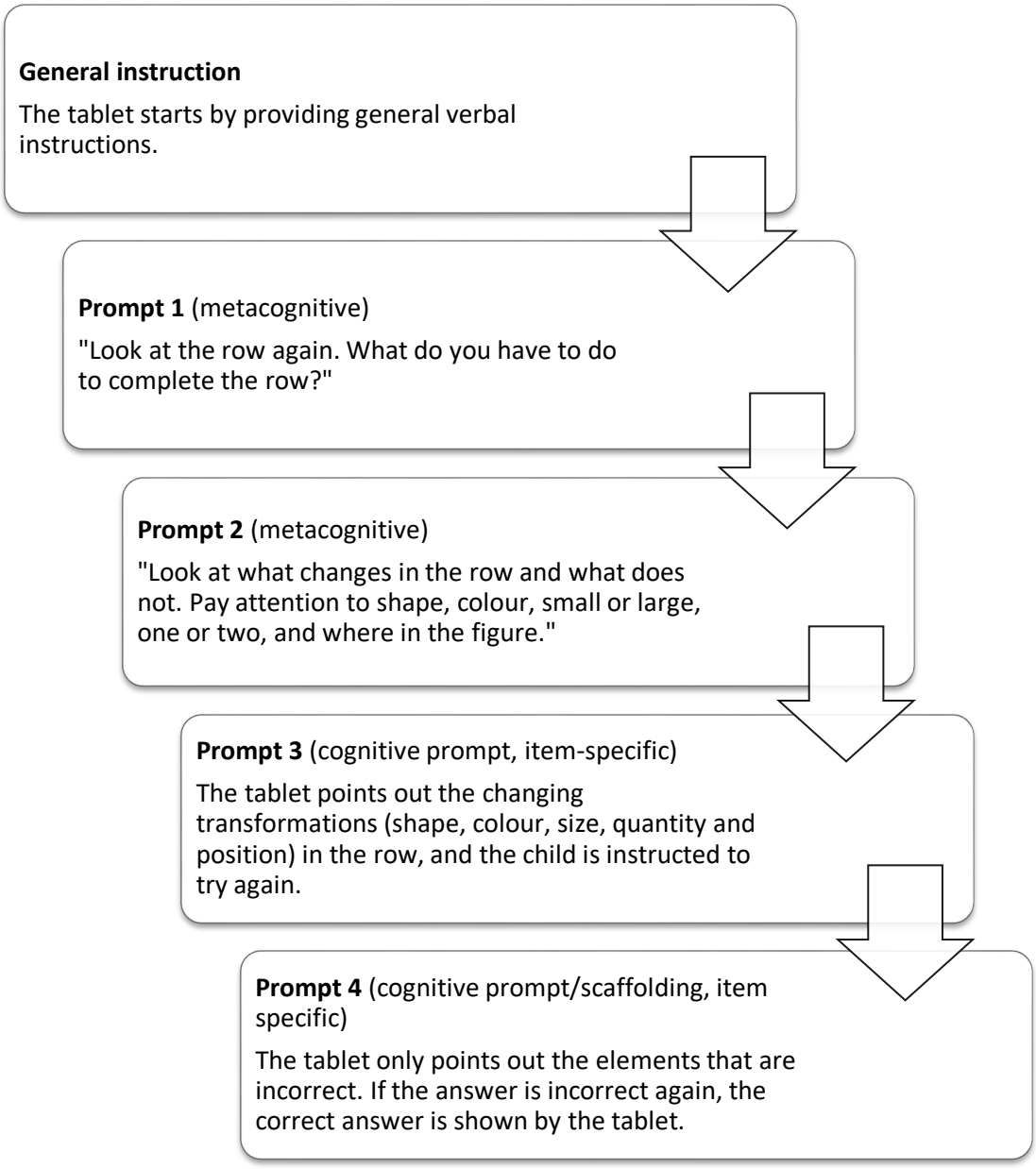

Figure 2. Schematic overview of the graduated prompts offered by the tablet during the dynamic training sessions

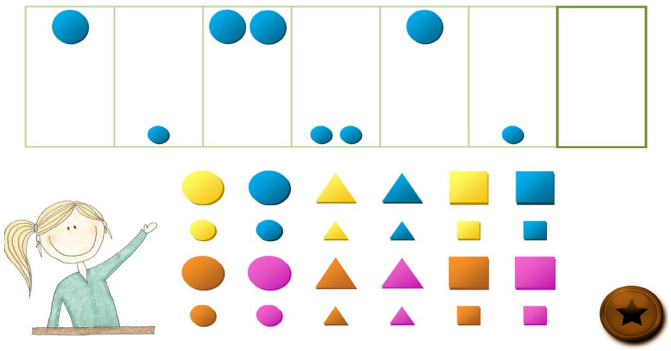

Figure 3. Display of the tablet with answering possibilities

To examine our second research question, the examiners assigned children's verbal explanations to one of 13 strategy categories, which are depicted in Table 2. These categories were separated into four main categories, partly on the basis of the categories used by Resing, Touw, and colleagues (2017): (1) no-answer: when no explanation or an unclear explanation is given; (2) non-inductive: when no inductive thinking is verbalized; (3) partial-inductive: when only one or a few (changing) transformations in the row are mentioned inductively; and (4) full-inductive: when an inductive description of all the changing transformations in the row is given.

To create strategy groups for each test session, a further categorisation was made: (1) no-answer; (2) mix of no-answer and non-inductive; (3) non-inductive; (4) mix of no-answer and partial-inductive; (5) mix of non-inductive and partial-inductive; (6) partial-inductive; (7) mix of partial-inductive and full-inductive; (8) full-inductive (see Table 2). Recordings of the verbal explanations of five children during the pre-test or post-test were not available; the data of these children were not included in the analysis. Interrater reliability was examined for the ratings of the verbal explanations of 70 children (44%) by calculating a two-way mixed-consistency-average intra-class correlation coefficient (ICC) per verbal explanation category. For the verbal explanation category 'no answer', ICC = .96 (95% CI = .94-.98); for the category 'non-inductive', ICC = .94 (95% CI = .90-.96); for the category 'partial-inductive', ICC = .97 (95% CI =.95-.98); and for the category 'full-inductive', ICC = .90 (95% CI = .83-.94).

Our third research question involved a tree analysis to determine interindividual differences in performance changes between the pre-test and post-test. We conducted a CRT tree analysis because it is the most suitable for data sets under N = 500 (Hayes, Usami, Jacobucci, & McArdle, 2015; Loh, 2009). Pruning was applied to avoid model overfit (Breiman, Friedman, Olshen, & Stone, 1984; Song & Lu, 2015; Wilkinson, 1992). We set 10 as the minimum number of cases in the parent node, and five was used as the minimum for each child node. We entered the following variables to investigate the influence on performance change: initial ability (pre-test score), condition, visual and auditory working memory, gender, and age.

Table 2. Verbal explanation categories and strategy groups

Category	Verbal explanation	Description
No-answer	Unknown	Explanation is inaudible, or child gives explanation from which a strategy cannot be deducted
	Guessing	The child doesn't know how he/she solved the task or guessed the answer
Non-inductive	Missing piece	Child used a figure because it was not in the row yet
	Fairness	Child aimed at an equal distribution of figures in the row
	Skipping the gap	Child only looks at certain boxes in the row
	Wishful thinking	Child changes one of the figures in the row for him-/herself, to make his/her answer fitting
	Repetition random square	Child repeats random figure from the row
Partial-inductive	Repetition first square	Child repeats first figure from the row
	Simple repetition	Child tries to find the figure in the row that is the same as the figure in box 6 and repeats the figure that comes after this
	Incomplete complex repetition	Child looks back in the row per transformation, like in simple repetition, but does not mention all changing transformations
	Incomplete seriation	Child mentions the pattern, but does not mention all changing transformations
Full-inductive	Complete complex repetition	Child looks back in the row per transformation, like in simple repetition, and combines these transformations. Child mentions all changing transformations
	Complete seriation	The child follows the row for all changing transformations

Strategy group	Criterion
1 No-answer	No answer explanation was used in more than 33% of the items
2 Mix of no-answer– non-inductive	Both categories were used in more than two times 33% of the items
3 Non-inductive	Non-inductive explanation was used in more than 33% of the items
4 Mix of no-answer–partial-inductive	Both categories were used in more than two times 33% of the items
5 Mix of non-inductive–partial-inductive	Both categories were used in more than two times 33% of the items
6 Partial-inductive	Partial-inductive explanation was used in more than 33% of the items
7 Mix of partial-inductive–full-inductive	Both categories were used in more than two times 33% of the items
8 Full-inductive	Full-inductive explanation was used in more than 33% of the items

3.3 Results

Before analysing the research questions, the comparability of the two groups of children in the experimental and control condition, respectively, was examined. Analyses of variance (ANOVA), using age in months and Raven's Progressive Matrices test score as the dependent variables and condition as the independent variable, revealed no significant differences between the children in the two conditions regarding age ($F(1, 162) = 2.245$, $p = .136$), or initial level of inductive reasoning as measured with the Raven ($F(1, 162) = .510$, $p = .476$), which indicated that participants in both conditions were comparable on these baseline variables. Table 3 provides an overview of the basic statistics between the children in the two conditions.

Table 3. Basic statistics of the children in the two conditions (control and training)

			N	*M*	*SD*
Gender	Control	Boy	39		
		Girl	45		
	Training	Boy	36		
		Girl	44		
Age in months	Control		84	94.36	5.17
	Training		80	95.50	4.56
Raven raw scores	Control		84	33.37	8.94
	Training		80	34.31	7.90
IRT gain scores	Control		84	-.25	.32
	Training		80	.27	.52
AWMA Spatial Recall Processing Standard Score	Control		70	109.21	18.88
	Training		68	107.40	20.48
AWMA Listening Recall Processing Standard Score	Control		70	109.59	17.67
	Training		68	114.51	15.36

Accuracy in solving series-completion task-items

Our first research question concerned the effect of training on children's progression in accuracy on a series-completion test. We hypothesized that as an effect of training, children in the experimental condition would improve their serial reasoning performance more than the untrained children in the control group, as indicated by their gain scores. We used the MRMLC model to answer this question. The base model (M0) assumes the person variables to be random. For the first model (M1), we added the main effect of Session, which resulted in a significantly better model fit, $p < .001$. In the second model (M2), the correlation between sessions was added to test the individual differences that arose between the pre-test and post-test. This model again led to a significantly better fit for the data, $p < .001$. In the third model (M3), the effect of Condition was incorporated to analyse whether children in the experimental condition progressed significantly more in reasoning accuracy than the children in the control condition. Adding the effect of Condition also led to a significant improvement to the model's fit, $p < .001$, which indicates a significant effect of Condition on children's reasoning accuracy. Table 4 displays the models' statistics and AIC and BIC values, with lower values indicating a better model fit. In conclusion, the analysis outcomes revealed that the trained children, when compared with the children in the control condition, made more progression in accurately solving series-completion task items (see Figure 4).

Table 4. Statistics for the IRT analysis investigating the effect of training

	Df	AIC	BIC	Log likelihood	Deviance	Chi-square	*df*	Probability (*p*)
M0	19	5091.5	5218.5	-2526.8	5053.3			
M1	20	4993.2	5126.8	-2476.6	4953.2	100.33	2	< .001
M2	22	4970.4	5117.4	-2463.2	4926.4	26.79	2	< .001
M3	24	4915.1	5075.5	-2433.5	4867.1	59.31	2	< .001

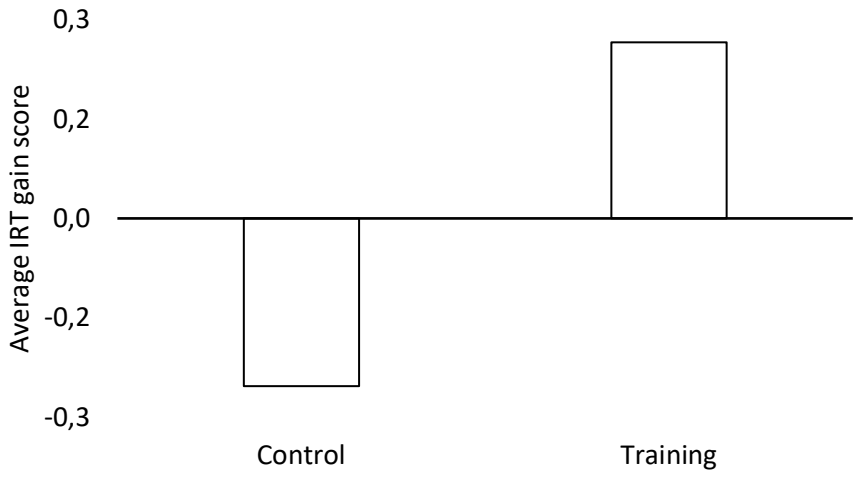

Figure 4. Schematic overview of the IRT gain scores

Verbal explanations

For our second research question, we examined the influence of two dynamic training sessions on children's verbal strategy-use. A multivariate repeated measures ANOVA was performed with Session (pre-test and post-test) as the within-subjects factor and with Condition (dynamic testing or control) as the between-subjects factor. The number of verbal explanations per strategy category (full-inductive, partial-inductive, non-inductive, and no answer) was used as dependent variables. Multivariate effects were found for the Verbal strategy category (Wilks' $\lambda = .062$, $F(3, 155) = 780.39$, $p < .001$, $\eta_p^2 = .94$), Session × Verbal strategy category (Wilks' $\lambda = .872$, $F(3, 155) = 7.56$, $p < .001$, $\eta_p^2 = .13$), Verbal strategy category × Condition (Wilks' $\lambda = .924$, $F(3, 155) = 4.23$, $p = .007$, $\eta_p^2 = .08$), and Session × Verbal strategy category × Condition (Wilks' $\lambda = .908$, $F(3, 155) = 5.25$, $p = .002$, $\eta_p^2 = .09$). The results of these analyses are depicted in Figure 5.

The univariate outcomes per verbal strategy category revealed no significant effects in both the *no-answer verbal strategy category* and the *partial-inductive verbal strategy category*. Training did not affect children's non-responsiveness or partial-inductive answers. Although the children who received training provided a larger number of partial-inductive verbal explanations, and the non-trained children at first sight showed a decrease in these explanations, these changes were not significant ($p = .107$). The analysis for the *non-inductive verbal strategy-category* revealed a significant interaction effect for Session × Condition: Wilks' $\lambda = .949$, $F(1, 157) = 8.51$, $p = .004$, $\eta_p^2 = .05$. Children in the control condition increased their non-inductive verbal explanations from the pre-test to post-test, and the children who received training showed a decrease of this non-advanced verbal strategy. In the *full-inductive verbal strategy category* significant main effects were found for Session (Wilks' $\lambda = .889$, $F(1, 157) = 19.66$, $p < .001$, $\eta_p^2 = .11$) and Condition ($F(1, 157) = 6.98$, $p = .009$, $\eta_p^2 = .04$), and a significant interaction was found for Session × Condition (Wilks' $\lambda = .964$, $F(1, 157) = 5.91$, $p = .016$, η_p^2

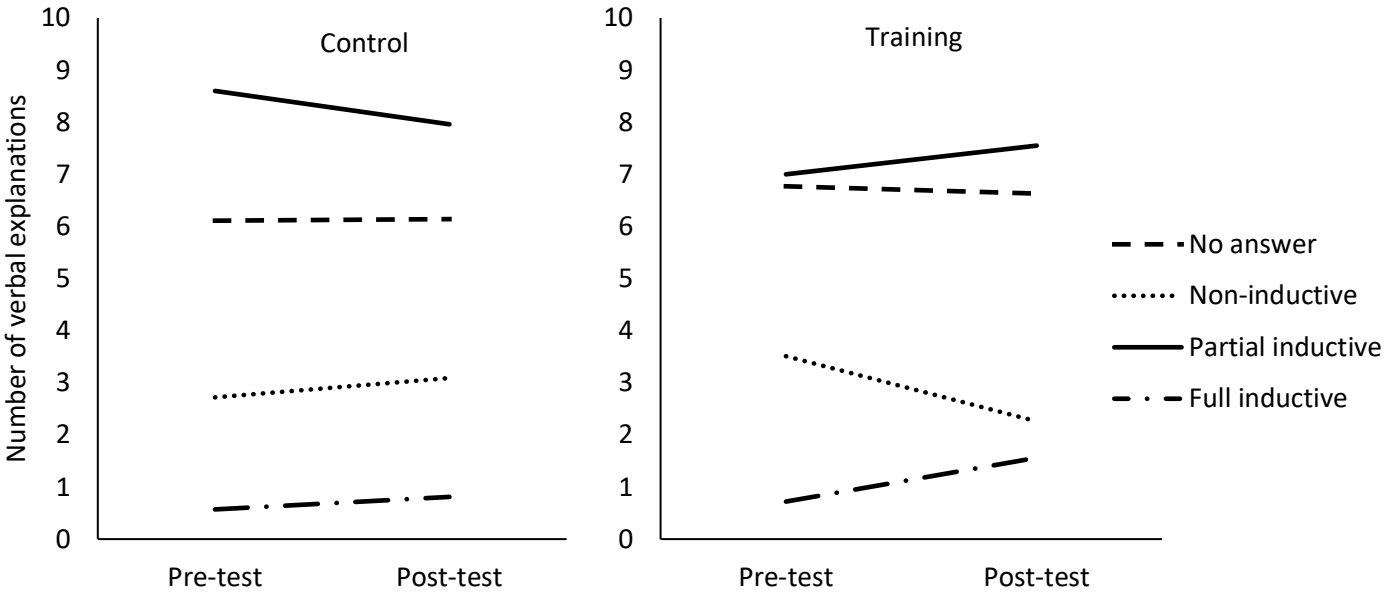

Figure 5. Patterns of change in verbal explanations of children in the training and control condition

= .04). Children used more advanced full-inductive verbal strategies in the post-test session, and training appeared to positively influence this progression.

To examine the effects of dynamic testing and verbal explanations, the children were assigned to different strategy groups. Crosstabs analyses (chi-squared tests) were used to investigate how children changed their verbal explanations over time. We examined shifts in verbal strategy use by analysing the relationship between Condition and Verbal strategy group (see Table 5). The pre-test results showed, as predicted, no significant association between the condition and types of verbalization (χ^2 pre-test (5, N = 153) = 6.80, p = .236, 33.3% of the cells had an expected count of less than 5). Unexpectedly, however, a non-significant association was found between the condition and verbal strategy- group for the post-test (χ^2 post-test (6, N = 156) = 7.38, p = .287, 28.6% of the cells had an expected count of less than 5).

Table 5. Change in verbal strategy groups from pre- to post-test, by condition

		Strategy group								
		1	2	3	4	5	6	7	8	Total
Pre-test										
Control	Frequency	19	2	3	6	6	43	0	0	79
	Percentage	24.1	2.5	3.8	7.6	7.6	54.4	0	0	100
Training	Frequency	25	3	7	4	8	27	0	0	74
	Percentage	33.8	4.1	9.5	5.4	10.8	36.5	0	0	100
Post-test										
Control	Frequency	22	0	9	10	3	35	0	0	79
	Percentage	27.8	0	11.4	12.7	3.8	44.3	0	0	100
Training	Frequency	25	0	5	5	8	32	1	1	77
	Percentage	32.5	0	6.5	6.5	10.4	41.6	1.3	1.3	100

Inter-individual changes in inductive reasoning

Our next research question concerned which factors influenced interindividual differences in gain scores between the pre-test and post-test of the computerized series-completion test. We used a tree analysis to answer this research question. Children's IRT-based gain scores were used as the dependent variable, while initial ability (pre-test score), condition, gender, age, standardized AWMA Listening Recall score, and standardized AWMA Spatial Span score were entered as predictors. Figure 6, showing the classification tree that resulted from the analysis, depicts each independent variable's contribution to the model. As Figure 6 shows, the condition is the first predictor that distinguishes children with large gain scores from those with small gain scores. Children in the training condition outperformed those in the control condition. Children in the training condition can be differentiated further by their initial ability: Children with a lower initial ability showed more improvement from the pre-test to post-test than children with a higher initial ability. The trained children with a higher initial ability can be differentiated further by their auditory working memory: Those with lower scores for their auditory working memory showed more improvement from the pre-test to post-test than the children with higher scores. Overall, condition and initial ability seem to be the most important predictors of children's progression in reasoning accuracy (see Table 6). Trained children with lower initial ability scores profited most from training.

Table 6. Independent variable importance to the model of change scores

Independent Variable	Importance	Normalized Importance
Condition	.067	100.0%
Total correct at pre-test	.025	37.6%
Age	.013	19.1%
AWMA Listening Recall Processing Standard Score	.009	13.0%
AWMA Spatial Span Processing Standard Score	.004	6.6%

3.4 Discussion

This study investigated children's progress in solving series completion after training by focusing on process-oriented assessment data captured by a tablet, including their reasoning accuracy and verbal explanations on a dynamic series-completion test. We compared the inductive reasoning progression between pre-test and post-test of children who received graduated prompts training with the progression of children who solved only the series-completion tasks twice without feedback. With IRT analysis, we were able to focus on gain scores of the individual children, which enabled us to conclude that children who received graduated prompts training achieved better learning gains in their series-completion skills than the children who received no training. These findings underline previous studies in which a dynamic testing approach has shown an additional effect of training on children's inductive reasoning accuracy (e.g., Resing & Elliott, 2011; Stevenson, Hickendorff, et al., 2013; Tzuriel & Egozi, 2010).

With regard to the verbal explanation strategies, our data revealed that children were categorized most often in the non-responsive and partial-inductive verbal explanations. However, the results did not show that training produced different strategy paths for these two verbal explanation categories. We did, however, find significant effects for the non-inductive and full-inductive verbal explanations, which children used less frequently. Children who received training utilized fewer non-

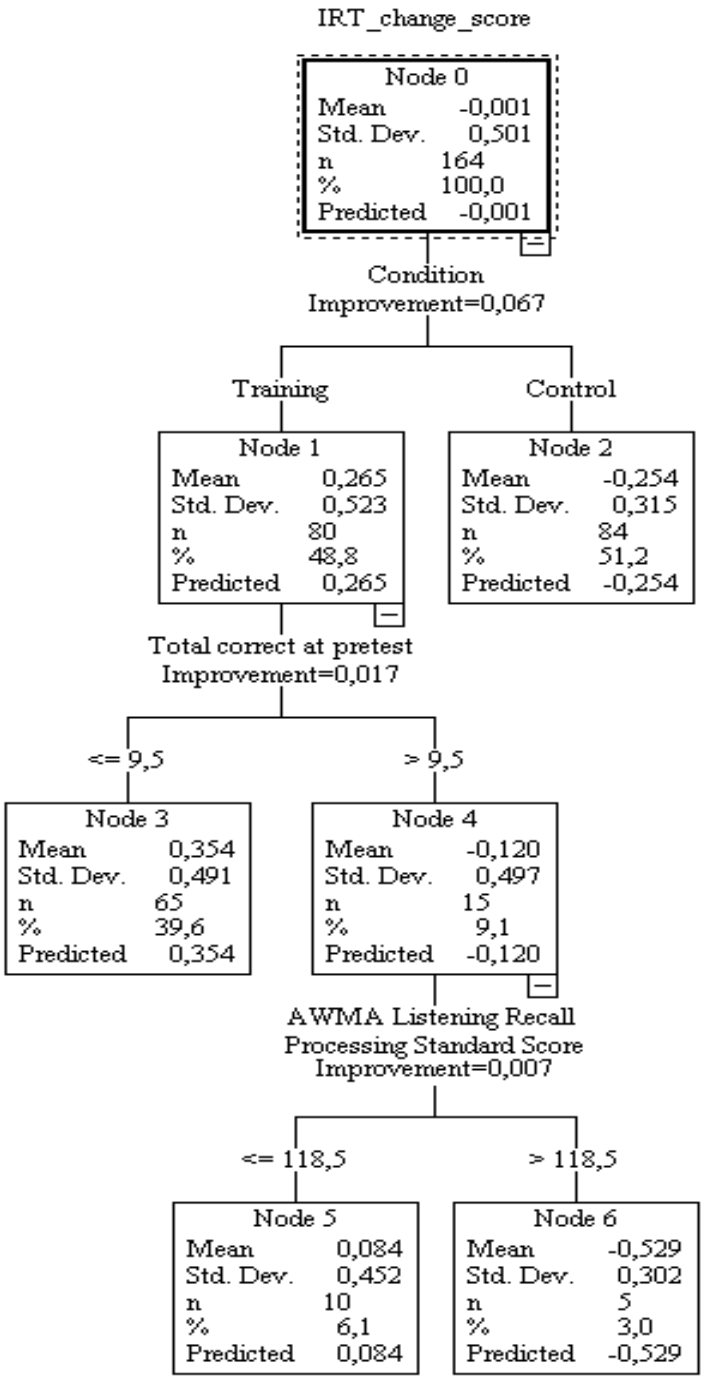

Figure 6. Classification tree of predictors (condition, pre-test scores, AMWA Listening Recall), influencing change scores

inductive verbal explanations and showed an increase in the advanced full-inductive verbal strategies in the post-test session. Our findings only partially support those reported by Resing, Bakker, and colleagues (2017), who found a strong increase of the advanced verbal strategy of inductive reasoning

after training was provided. Children's infrequent use of full-inductive verbal explanations in our study might have occurred because the children in the current study were younger, and our task appeals less to step-by-step task solutions, which may affect children's verbal explanations. The series-completion test used in this study asks for a more holistic approach to solving a global task when compared with, for example, the puppet task used by Resing and colleagues (2015) and Resing, Bakker, and colleagues (2017). Moreover, when the children were asked to explain their answers, the question did not clearly indicate that they should name as many transformations as possible. Since the dynamic test we constructed was made less verbal than tests developed before, no explicit training in verbally explaining their answers was provided, and though the transformations were mentioned and modelled in the training, verbalizing them was not the primary purpose of the training.

Another aspect of the current study that should also be considered in future studies on children's verbal explanations is the difficulty level of the task items. It might be worthwhile to examine verbal explanations for the easy and difficult items separately because more full-inductive answers would be expected for the easy items, as these items comprise fewer transformations.

When studying children's ability to change, in relation to strategy use, we examined both their development in verbal explanations and in overt problem-solving behaviour, as posited by Siegler and Svetina (2006). However, verbal explanations might not always be reliable indicators of children's problem-solving processes, especially for those as young as 7 to 8 years old (Resing et al., 2012). Including children's detailed problem-solving, for example, their overt problem-steps, behaviour would potentially provide more insight into individual differences of children's problem-solving processes. Future studies on dynamic testing and the development of children's strategies should consider both aspects.

In addition to children's development in accurately solving and explaining series-completion tasks, we were interested in the factors that influence individual differences in solving series-completion tasks. Our results showed that receiving training and children's initial ability were the most important predictors of children's increase in reasoning accuracy. Trained children, especially those who had a lower initial ability, outperformed untrained children. Also, trained children with a higher initial ability plus a relatively lower auditory working memory showed more improvement from the pre-test to post-test than did the children with higher scores for their auditory working memory. These results highlight the importance of dynamic testing for children with weaker initial reasoning skills or auditory working memories. Computerized dynamic tests, such as the one utilized in this study, certainly generate more information regarding the process of solving tasks individual children show. The assessment outcomes, reported by educational or school psychologists, reveal what children do with the feedback provided during dynamic testing and could influence teachers' views on how individual children could be supported in their learning, thereby contributing to formative assessment.

Computerized dynamic testing is a promising starting point for designing an efficient, integrated, and student-centred learning environment. Whether teachers can easily implement these assessment outcomes in their teaching and educational plans will have to be a focus of study in the future (e.g., Bosma, Stevenson, & Resing, 2017). Moreover, the benefits of dynamic testing lie in the fact that this method aims to focus on individual needs and can be seen as a potentially useful addition to conventional static tests used to predict school achievement (Caffrey et al., 2008; Fabio, 2005). Such predictions are important as they can identify students at risk of school failure as well as those in need of a more intensive intervention (Caffrey, Fuchs & Fuchs, 2008; Resing & Drenth, 2007). As part of the current study, no scholastic achievement data of children were collected. Therefore, the predictive value of dynamic tests in relation to scholastic achievement needs to be a focus point of future studies (e.g., Jeltova et al., 2007; Yang et al., 2017). Some overall limitations of the dynamic series-completion test used in the current study included that the training approach consisted of two short training sessions and no follow-up after the post-test. Because children were tested during school hours, it was not possible to increase the length of the training sessions. In future studies, however, it would be worthwhile to investigate whether a more intensive training procedure, for instance one that contains more items or a larger number of training sessions, would lead to different progression paths in the context of accuracy and children's verbal explanations, as well as larger interindividual differences. Moreover, future studies could implement a follow-up session to investigate to what extent children retain the skills and knowledge acquired as part of the dynamic test.

Furthermore, the technological possibilities of using a tablet should be explored further. For example, we did not program the tablet to record children's verbal explanations. The test examiner used a separate voice recorder, which was an extra action for the examiner and more time-consuming. The benefits of using electronic technology in the field of dynamic testing are numerous, and computer technology can create new methods for examining problem-solving processes in more depth (Resing & Elliott, 2011; Tzuriel & Shamir, 2002). Computerized testing can provide additional information that may be useful for individualized (educational) instructions, problem-solving processes, and intervention (Passig et al., 2016; Resing & Elliott, 2011; Stevenson et al., 2011).

The current study has shown that providing children a dynamic graduated prompts training leads to a positive change in their reasoning abilities in a series-completion test. More information was obtained about the cognitive-development trajectories of children, providing us with better understanding of how learning occurs and which factors contribute to cognitive change. Because static testing can lead to the underestimation of children's actual cognitive level, future research should focus on more process-oriented assessment techniques, such as dynamic testing. In doing so, the dynamic test of series completion utilized in the current study could be employed to assess children's reasoning ability, as series completion is a subform of inductive reasoning, as a measure of their fluid

intelligence. As the test items are constructed using geometric shapes, it can be argued these are relatively culturally non-sensitive, being appropriate for testing children of diverse cultural and linguistic backgrounds. Of course, for these target groups the verbal instructions provided may need to be adapted. These aspects will be valuable topics for future research, investigating the wider applicability of the dynamic test utilized in the current study.

Advances in computerized dynamic testing may establish testing methods that can provide both adaptive and standardized means of examining children's problem-solving processes and the development of their cognitive abilities. Implementation of the assessment outcomes in classroom learning and thereby enhancing learning opportunities in children have to be studied in the future (e.g., Stringer, 2018). Computerized dynamic testing can be considered a good step in that direction.

Appendix A

Detailed overview of the computerized series-completion test presented on the tablet

Display		The tablet display visually presents the task to the child. A puppet named Lisa provides instructions and prompts or feedback when necessary. The row with geometric figures is shown on the display of the tablet. The child can tap on a basket to reveal the geometric shapes in four different colours and two sizes (see Figure 1), select the shape he or she wants to use, and drag it to the empty box in the row of figures. When the child drags a shape into the last (empty) box of the figure he or she needs to press a star-button in order to confirm the answer and proceed to the next item (see Figure 1)
Instructions	Pre- and post-test	The tablet provides general verbal instructions during the pre- and post-test. At the beginning of the pre-test, Lisa introduces herself and explains all the shapes, colours and sizes of the geometric figures that are used in the test, to familiarize the children with the figures and corresponding names. For both the pre- and post-test, two example items are given in order to explain the task. Lisa describes the figures in the row. Then, the child is asked to complete the row. If the child provides an incorrect answer, the correct answer is explained. If the answer is correct, the child receives positive feedback. After the example items, feedback is not provided anymore. After each item, the child is asked why they chose their answer.
	Training	The general instructions in the training procedure are similar to the instructions in the pre- and post-test. After each correct answer, the child receives positive feedback and is asked why they chose their answer. When an answer is incorrect, prompts are provided by the tablet.
Audio effects		The tablet provides additional auditory feedback after an answer is given during the example items and the training procedure. A high 'pling' sound is played whenever an answer is correct and a lower sound when the child's answer is incorrect.
Visual effects		The tablet provides visual effects parallel to the verbal instructions to visually attract attention to the figures. The tablet briefly enlarges the geometric figures in the row, the outlines of the boxes and the outline of the complete row.