



Universiteit
Leiden
The Netherlands

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

Touw, K.W.J.

Citation

Touw, K. W. J. (2020, September 17). *Computerised dynamic testing: An assessment approach that tailors to children's instructional needs*. Retrieved from <https://hdl.handle.net/1887/136755>

Version: Publisher's Version

License: [Licence agreement concerning inclusion of doctoral thesis in the Institutional Repository of the University of Leiden](#)

Downloaded from: <https://hdl.handle.net/1887/136755>

Note: To cite this publication please use the final published version (if applicable).

Cover Page



Universiteit Leiden



The handle <http://hdl.handle.net/1887/136755> holds various files of this Leiden University dissertation.

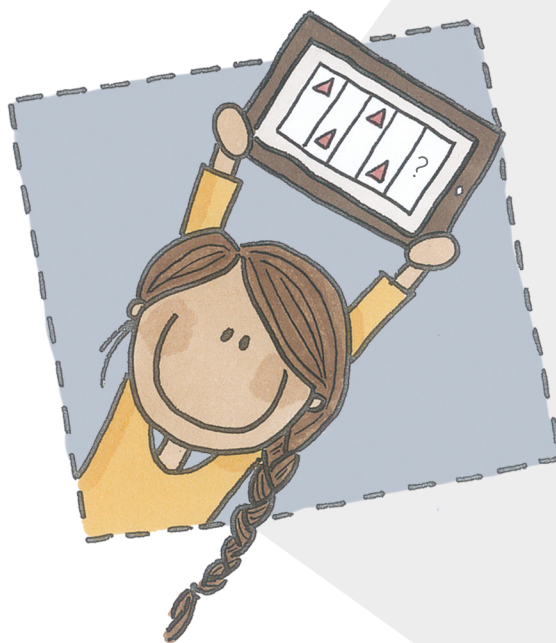
Author: Touw, K.W.J.

Title: Computerised Dynamic Testing: An assessment approach that tailors to children's instructional needs

Issue Date: 2020-09-17

Chapter 2

USING ELECTRONIC TECHNOLOGY IN THE DYNAMIC TESTING OF YOUNG PRIMARY SCHOOL CHILDREN: PREDICTING SCHOOL ACHIEVEMENT



KIRSTEN W. J. TOUW
BART VOGELAAR
MEREL BAKKER
WILMA C. M. RESING

Touw, K. W. J., Vogelaar, B., Bakker, M., & Resing, W. C. M. (2019). Using electronic technology in the dynamic testing of young primary school children: Predicting school achievement. *Educational Technology Research and Development*, 67(2), 443-465. <https://doi.org/10.1007/s11423-019-09655-6>

Abstract

This study aimed to combine the use of electronic technology and dynamic testing to overcome the limitations of conventional static testing, and adapt more closely to children's individual needs. We investigated the effects of a newly developed computerized series completion test using a dynamic testing approach and its relation to school achievement. The study utilized a pre-test-training post-test control-group design in which 164 children from grade 2 participated. To evaluate the additional effects of dynamic testing beyond the effects of (repeated) static testing of inductive reasoning on a tablet, half of the children were trained using a graduated prompts method, while the other half of the children only practiced solving the series completion task-items. The results showed that training with graduated prompts is effective in increasing the likelihood that children can solve series completion problems accurately. Furthermore, the number of prompts children needed during training, significantly predicted the performances of children on mathematics and technical reading tests. Teacher's judgments regarding their pupils' overall school performance and potential for learning, however, did not correlate significantly with the dynamic post-test score of the series completion test, which seemed to indicate that dynamic testing provides teachers with new information about the learning progress of individuals.

2.1 Introduction

In contemporary education, teachers often make use of interactive boards, video support, tablets, and mobile technology that have been developed to assist their teaching (Pamuk, Çakır, Ergun, Yılmaz, & Ayas, 2013; Föbl, Ebner, Schön, & Holzinger, 2016). In general, using computerized instructional designs supports the provision of immediate and individualized feedback to the child. Moreover, computerized help or instruction may result in creating a more authentic assessment environment (Huang, Wu, Chu, & Hwang, 2008; Khandelwal, 2006), including more systematic instruction or assessment procedures (Tzuriel & Shamir, 2002). The increasing use of technology in schools is often said to support children in solving school tasks (Haßler, Major, & Hennessy, 2016), and cater to their individual needs.

The focus on students' individual needs is also embedded in a worldwide trend of introducing student-centered educational systems, in which reasoning and problem solving are encouraged, and individuals are enabled to address unique learning interests and needs (Azevedo, Behnagh, Duffy, Harley, & Trevors, 2012; Hannafin & Land, 1997). This trend leads to a growing focus on the (assessment of) abilities and educational needs of individual students. Linked in with this, there is a need for assessment procedures that take into account individual differences, and provide an indication of children's individual instructional and further educational needs. Dynamic testing, using a test-training-test format, comprises one way of shedding more light on students' individual needs. Therefore, the aim of our study was to zoom in on computerized assessment and feedback in a dynamic testing setting, and explore the effects of a newly developed computerized series completion test.

Computerized dynamic testing

Unlike conventional forms of static testing, where by convention no feedback on how to solve tasks and improve performance is given, dynamic testing incorporates feedback and training into the testing phases, thereby providing information about the individual's progression in performance and cognitive functioning (Grigorenko, 2009; Haywood & Lidz, 2007; Resing, 2013; Resing, Elliott, & Grigorenko, 2012). Whereas conventional tests to measure children's cognitive abilities, such as standardized intelligence tests, for a large part rely on previous learning experiences, dynamic testing focuses on the individual's potential for growth and learning abilities (Grigorenko & Sternberg, 1998), a notion which is derived from Vygotsky's theory of the Zone of Proximal Development (Vygotsky, 1978). Studies have consistently shown that dynamic testing is an effective method to gain more insight into children's individual cognitive strengths and needs (e.g., Bosma, Stevenson, & Resing, 2017; Caffrey, Fuchs, & Fuchs, 2008; Elliott, 2003; Hill, 2015; Jeltova et al., 2007; Resing & Elliott, 2011; Resing, Elliott et al., 2012; Tzuriel, 2000).

Although dynamic testing has clear advantages to traditional static testing, it typically includes several assessment points, which is why administering these tests can be time-consuming, and therefore, rather difficult to apply in practice. As a result, the use of computerized dynamic testing is gaining more and more attention. Computerized testing procedures enable registering scores automatically, and show results immediately after testing, which is more time-efficient and leaves less room for errors in administering and interpreting test outcomes. Several researchers have shown the positive value of incorporating computer-assisted feedback during a dynamic test (e.g., Passig, Tzuriel, & Eshel-Kedmi, 2016; Poehner & Lantolf, 2013; Resing & Elliott, 2011; Resing, Steijn, Xenidou-Dervou, Stevenson, & Elliott, 2011; Stevenson, Touw, & Resing, 2011; Tzuriel, & Shamir, 2002). For example, Resing and Elliott (2011) developed a highly structured visual-spatial puppet series completion test using tangibles on an electronic console. While the use of multiple tangibles (blocks with electronic identification codes) in constructing the test items provided detailed information on the solving processes of individual children, administering such a computerized test is still quite time-consuming as the child needs to provide an answer by manipulating eight blocks that have to be placed back on the table in an orderly way, after solving each task item.

Therefore, in the current study, we constructed a new dynamic series completion test using a whole-figure geometric solution instead of a multiple-parts item solution, for 6-8 year old children, as the first few years of primary school are considered to be a period of rapid development of the ability to reason inductively (Siegler & Svetina, 2002). Solving this kind of tasks follows two levels of inductive reasoning: both the number of pattern changes in a series and the period of change (periodicity) have to be detected to formulate a solving rule (Resing & Elliott, 2011). We administered this dynamic series completion test on a tablet instead of on an electronic console or a computer, because we assumed a test on a tablet could be easily incorporated into the classroom and handled by young school children. In addition, studies have demonstrated that providing instruction on a mobile device such as a tablet increases children's motivation beyond the effect of traditional classroom instruction (Furio, Juan, Segui, & Vivo, 2015), which some authors attribute to challenge, immediate feedback, the sense of control, recognition, competition and cooperation that are enabled by using mobile devices (Ciampa, 2013). Therefore, the tablet can be seen as an authentic assessment instrument, in a seamless-learning-setting (Föböl et al., 2016; Schmitz, Klemke, Walhout, & Specht, 2015) for young school children when solving series completion tasks during dynamic testing.

Dynamic testing in relation with school achievement and teachers' judgments

Different from static testing, dynamic testing 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 et al., 2008; Resing

& Drenth, 2007). Past studies uncovered a clear relationship between (dynamic) tests of inductive reasoning and school achievement measures (e.g., Beckmann, 2006; Csapó, 1997; Fabio, 2005; Klauer & Phye, 2008; Resing, 1993; Sonntag, 2006; Stevenson, Bergwerff, Heiser, & Resing, 2014; Tzuriel & George, 2009). More specifically, a relation has been reported between scores on series completion tests and achievement test scores related to specific school subjects, such as reading and writing (Ricketts, Bishop, & Nation, 2009) and mathematics (Primi, Ferrob, & Almeida, 2010; Stevenson, Hickendorf, Resing, Heiser, & De Boeck, 2013; Taub, Floyd, Keith, & McGrew, 2008; White, Alexander, & Daugherty, 1998). In their review of fifteen studies into static and dynamic assessment, Caffrey and colleagues (2008) found an average correlation of $r = .49$ between dynamic assessment scores and overall achievement measures of preschool students. The average correlation between traditional static testing and overall achievement measures of preschool students was $r = .41$. Although both correlations are not significantly different from each other, dynamic testing procedures focus on change processes, and, as a result, potentially provide more information than static tests.

Previous studies have also addressed the relationship between teachers' judgments of students' academic achievements and performance measures of their academic achievements (Feinberg & Shapiro, 2009; Hoge & Coladarci, 1989; Südkamp, Kaiser, & Möller, 2012). These authors have repeatedly sketched moderate to strong correlations between teachers' judgments and students' school achievement variables between grade 2 and 5 in relation to reading, arithmetic skills and overall academic achievement. Only few researchers have reported a link between dynamic testing outcomes and measures of teachers' judgment (Bosma & Resing, 2008; Resing, 1993). In these studies, it was found that dynamic tests provided an additional predictive value in relation to school performance and teacher ratings, which was stronger than that of static tests. More importantly, Bosma and Resing (2008) studied teachers' evaluations of diagnostic reports on the potential for learning of the children. In this study, it was found that teachers' judgment scores concerning their pupils' potential for learning were significantly related to their pupils' scores on a dynamic test ($r = -.61$).

Aims of the current study

The main aim of the current study was to examine the potential effects on children's progression in series completion of a newly developed dynamic test, consisting of computerized geometric series completion task-items. Implementing this dynamic test on a tablet would, in principle, make it more consistent with the increasing use of technology in the classroom. By assessing the dynamic test on a tablet we could effectively interact with the children and efficiently record their performance. Secondly, we examined to what extent outcomes of the dynamic series completion test were related to children's school performance and teachers' ratings of their school performance and cognitive capabilities.

In our study we focused on the following underlying issues. The first research question concerned the potential differential effects on the post-test of the children in the two conditions (training versus control group) brought about by the computerized graduated prompts procedure. Based on earlier findings, it was hypothesized that trained children would improve their task solving accuracy, defined as the number of (a) correct solutions and (b) correctly applied task-transformations, more than control-group children (e.g., Campione, Brown, Ferrara, Jones, & Steinberg, 1985; Resing & Elliott, 2011; Resing, Xenidou-Dervou et al., 2012).

The second research question focused on the relationship between dynamic testing outcomes and school achievement measures. We hypothesized that the dynamic measures of the series completion test would hold predictive value in relation to school achievement measures that was at least equally strong as the static series completion measures (e.g., Caffrey et al., 2008). In earlier studies, reporting on differences in the relationship between either dynamic or static measures and school achievement, these differences were usually (modestly) in favor of the dynamic measures (e.g., Beckmann, 2006; Fabio, 2005; Resing, 1993; Stevenson et al., 2014).

The third and final research questions concerned the relationship between teachers' judgments of children's scholastic performance and children's potential for learning. Findings of earlier studies showed that teachers' judgments regarding children's scholastic achievements were fairly accurate (Bosma & Resing, 2008; Feinberg & Shapiro, 2009; Hoge & Coladarci, 1989; Resing, 1993; Südkamp et al., 2012). Based on earlier research findings, we hypothesized that teacher ratings of children's school performance and their potential for learning would be more strongly related to children's dynamic than to their static series completion test measures, because dynamic test outcomes have been defined as reflecting learning as a result of the extensive instructions during the training phase (e.g., Resing, 2013).

2.2 Method

Participants

The study employed 164 children (89 girls and 75 boys) with a mean age of 94.91 months ($SD = 4.9$ months). The children, recruited from fourteen primary schools, attended classes at a second grade. These schools were located in midsize and large towns in the western part of the Netherlands. Children's primary language spoken at school was Dutch. For all children, written school and parental informed consent for participation was obtained in two steps. First, a random selection of schools in the proximity 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 parental consent for participation was obtained for all children. Initially, the study

included 177 children. Thirteen children, however, dropped out of the study during the extended period of testing, due to absence or illness during one or more test sessions. No further exclusion criteria were applied. In total, 17 schools and 17 teachers participated in the study. All procedures, including the informed consent and the recruitment of participants, were reviewed and approved by the institutional Committee Ethics in Psychology (CEP).

Design and procedure

An experimental pre-test-training-post-test control-group design was employed (see Table 1). The design shows a combination of a static pre-test measuring the current abilities, and a dynamic post-test measuring a change in abilities following training (e.g., Day, Engelhardt, Maxwell, & Bolig, 1997; Resing, 2000). To reduce differences in initial inductive reasoning abilities between conditions, per school, we used randomized blocking based on children's initial inductive reasoning ability, measured by the Raven's Progressive Matrices test (Raven, Raven, & Court, 1998). Per school, children were paired based on their Raven scores and then randomly allocated to either the training or the control condition. Each child in the study took part in five individual weekly sessions (see Table 1) with each session lasting approximately 30 minutes. Children in both conditions completed a pre and post-test without receiving feedback as to the correctness of their answers. The treatment of children in the training condition differed from those in the control condition, because the first group of children received a two-session training between the pre-test and post-test, whereas the control-group children solved paper-and-pencil control tasks (mazes and dot-to-dot completion tasks), taking approximately the same time as the two dynamic training sessions, in order for the contact moments with the test leader to be as equal as possible between the two groups. All tests were administered individually in a quiet room in the child's school. Examiners, seated next to the child, were 10 well-trained master students who followed courses in educational and child psychology. After ending the study, schools were provided with a short report of the anonymized study outcomes.

Table 1. Schematic overview of the design of the study

Condition	<i>N</i>	Raven (session 1)	Pre-test (session 2)	Training 1 (session 3)	Training 2 (session 4)	Post-test (session 5)
Training	80	Yes	Yes	Yes	Yes	Yes
Control	84	Yes	Yes	No/mazes	No/mazes	Yes

Materials

Raven's Progressive Matrices. Raven's Progressive Matrices Test (Raven et al., 1998) is a non-verbal test of fluid intelligence, consisting of 60 multiple-choice items, referring to children's inductive reasoning ability and problem solving skills. Children were asked to complete a figure by choosing the

missing element. The Raven test has a reliability of $\alpha = .83$ and a split-half coefficient of $r = .91$ (Raven, 1981).

Measures of school achievement. Among primary school children in the Netherlands, academic achievement is commonly assessed by standardized school achievement tests developed by the Dutch institute for test development (Cito). These tests are part of a monitoring and evaluation system, and are administered twice a year to provide an indication of the child's performance in several academic domains compared with the national norms per age group. Cito scores range from A–E, with A being very good and E being very poor: A=25% highest scoring children, B=25% (well to just) above average scoring children, C=25% (well to just) below average scoring children, D=15% well below average scoring children, and E=10% lowest scoring children (Hollenberg, Van der Lubbe, & Sanders, 2017). For this study, Cito scores were converted into numeric, ordinal scores ranging from 1 (lowest) to 5 (highest), and therefore ordinal regression analyses were performed to study the predictive value of the series completion test in relation to children's school achievement. The 'Mathematics', 'Technical Reading / DMT' and 'Spelling' Cito test results were provided by the participating schools.

Cito Mathematics (M4 [grade 2]). The main objective of Cito Mathematics (Janssen, Hop, & Wouda, 2015) is to provide an indication of the child's level of mathematical ability. The test measures, among others, counting, classifying, dividing of numbers, addition, subtraction, multiplying and automatism. The reliability, defined in terms of measurement accuracy, is $MAcc = .93$ (Janssen et al., 2015).

Cito Technical Reading (M4 [grade 2]). The Three Minutes Test (DMT, Krom, Jongen, Verhelst, Kamphuis, & Kleintjes, 2010) measures the child's accuracy in and speed of reading individual words aloud, and gives an indication of the technical reading ability of the child. Cito DMT has a reliability of $\alpha = .97$ (Krom et al., 2010).

Cito Spelling (M4 [grade 2]). Cito Spelling (Tomesen, Wouda, Mols, & Horsels, 2015) provides an indication of the child's level and development of spelling skills. The test consists of one and two-word, and sentence dictations. The reliability, defined in terms of measurement accuracy, is $MAcc = .89$ (Tomesen et al., 2015).

Teacher ratings. The teachers were asked to complete a rating form, in which they placed the child's school performance in each school subject on a 5-point scale by comparing the child with their age-mates. Additionally, two questionnaire items were used: (1) How do you (the child's teacher) score the level of the child's learning ability in comparison to his or her age-mates? (rating the teachers' judgment as to the potential for learning of the child); and (2) What is your (the child's teacher) overall impression of the child's school performance in comparison to his or her age-mates? (rating the teachers' judgment as to the overall school performance of the child). A score of 1 placed the child among the 20% lowest performing; a score of 5 among the 20% highest performing children. The test

leaders provided the teachers with the paper-and-pencil questionnaire. General instructions on how to complete the ratings could be found on the questionnaire, and the teachers could ask the test leader any additional questions if necessary.

Dynamic test of series completion

Test construction. For this study, a new computerized visual-spatial series completion test was constructed. This test required children to detect rules and regularities by inductive reasoning, and the task construction was based on earlier developed series completion tasks: schematic-picture puppet series tasks (Resing & Elliott, 2011; Sternberg & Gardner, 1983) and letter series completion tasks (Ferrara, Brown, & Campione, 1986; Simon & Kotovsky, 1963). In comparison with series completion tasks including letters or numbers, pictorial and geometrical series completion tasks are more complex, because the elements in the series do not have a fixed, predictable, relation to each other (Quereshi & Seitz, 1993; Resing & Elliott, 2011). Moreover, in these tasks, children are less likely to be influenced by language or task-specific knowledge (Hosenfeld, Van den Boom, & Resing, 1997).

In the newly constructed series completion test, we used series each consisting of a row of six boxes filled with geometric figures, and a seventh, empty box. The series each included configurations of these figures based on three different geometrical figures: circles, squares and triangles. The series could include one to five different transformations: changes in geometrical shape, color, size, quantity, and position. Solving the test items therefore required task solving on two levels: the number of pattern transformations and the period of change (periodicity) (Resing et al., 2015; Simon & Kotovsky, 1963). Discovering periodicity involves noticing that patterns are repeated at predictable, regular intervals (Holzman, Pellegrino, & Glazer, 1983).

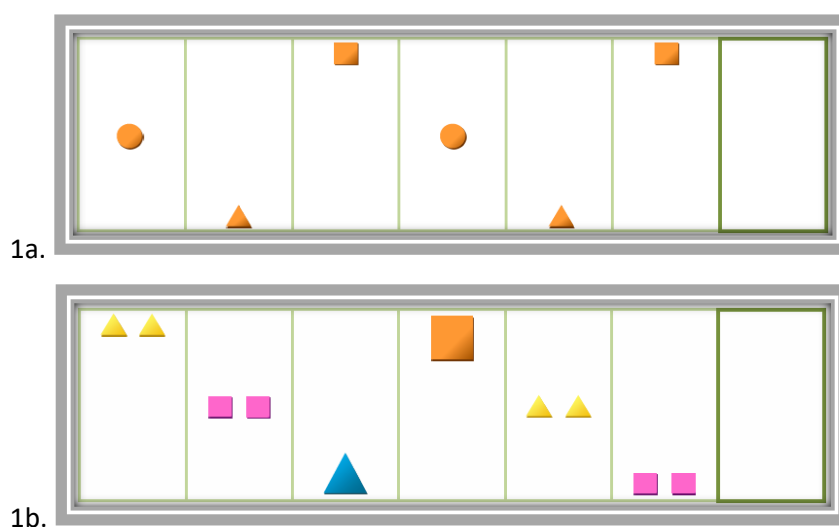


Figure 1. Geometric series completion test. (a) easy item with 2 transformations; shape and position (with periodicity 3 - abcabc). (b) difficult item with 5 transformations; shape (periodicity 2), color (periodicity 4), size (periodicity 2), quantity (periodicity 2), and position (periodicity 3).

The children needed to determine the changes in the row pattern and find the correct solution by discovering the underlying rule(s). They were asked to complete each row by placing one or multiple figures in the seventh box, in the correct position. Figure 1 provides an example of a series completion task-item. Children could solve a series completion item by pushing the virtual button on the tablet screen, after which all the geometric figures needed to complete a series were shown. They had to drag and place one or more of these figures into the right position in the empty box (see Figure 2). Throughout the test and training sessions, after each item the children were asked to explain verbally why they thought the solution they chose was the correct one. Due to the fact that children were being tested during school hours, they could be asked to complete a limited number of items only. All items used in the current test were piloted by the developing team with children of the same age to examine their suitability for this age group. Administration of the dynamic test of series completion was conducted by 10 master's students in educational and child psychology.

Both pre- and post-test consisted of 18 geometric series to complete, with two additional example items which were provided to familiarize children with the task demands. As the number of transformations and the periodicity of the items increased, so did the difficulty level (see Appendix A). The items were presented on a tablet (see Figure 2). The pre and post-test were parallel versions, with different but isomorphic items, and exactly the same testing procedure.

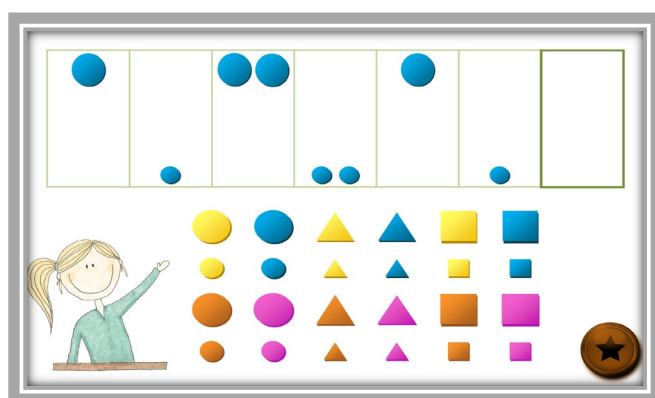


Figure 2. Display of the tablet with answering possibilities. By pushing the virtual button (with the star), all possible answering figures are shown, and the child must drag one or more of the geometric figures to the right position in the empty box.

Training procedure. The training consisted of two training sessions. In each session, the children were given six series completion items comparable to the ones they received during the pre and post-test. Standardized assistance was provided using a graduated prompts procedure (e.g., Campione & Brown, 1987; Ferrara et al., 1986; Resing, 1997; Resing & Elliott, 2011), which consisted of predetermined prompts that range from general to specific instruction. A new prompt was only

provided when the child's response was inaccurate. The order of the items presented during the training sessions ranged from difficult to easier ones. If the 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 emphasizing the specific transformations in the row was provided. If the child was still unable to accurately solve the task, direct guidance by scaffolding was provided. No further prompts were provided when a child solved an item accurately. Children were provided with visual and oral information and sounds (see Figure 3). An overview of the training procedure, including a screenshot of the different visual prompts provided, can be found in Appendix B.

Electronic device: Tablet

The series completion test was created using GM:Studio (Yoyogames). The items were run on an Acer Aspire Switch 10 convertible tablet, which had a Windows operating system and a 10.1 inch touch screen display with a resolution of 1280x800 pixels. The tablet was programmed to present the series completion items using different forms of output (display, verbal instructions, auditory and visual feedback). See Figure 3 for a detailed overview of the programming of the computerized series completion test presented on the tablet.

The tablet was also programmed to automatically score children's performance during the pre-test, training and post-test by using log files. For each of the 18 pre and post-test items, answers were scored as accurate (1) or inaccurate (0). Composite scores were computed for the total number of accurately solved items (range 0-18). The number of correctly applied transformations in each item could range from one to five. The six items per training session were scored similarly to the pre- and post-test scores. We used repeated measures analyses of variance (RM-ANOVA) to investigate the effect of the dynamic series completion test by looking at children's improvements in inductive reasoning (i.e., accurate solutions and number of transformations correct).

2.3 Results

Task characteristics

The series completion test included items of different difficulty levels. The expected theoretical difficulty of the items was based on the number of transformations in a row and on the frequency of recurring periods of change (periodicity), based on the model of Simon and Kotovsky (1963). We examined whether the construction of the items based on this theoretical model corresponded to the empirically measured *p*-values (the percentage of the children who answered the item accurate) of the pre-test. Appendix A provides these *p*-values. We expected that increasing the number and periodicity of changes would result in an increase in the difficulty level, and, as a result, in the number of errors

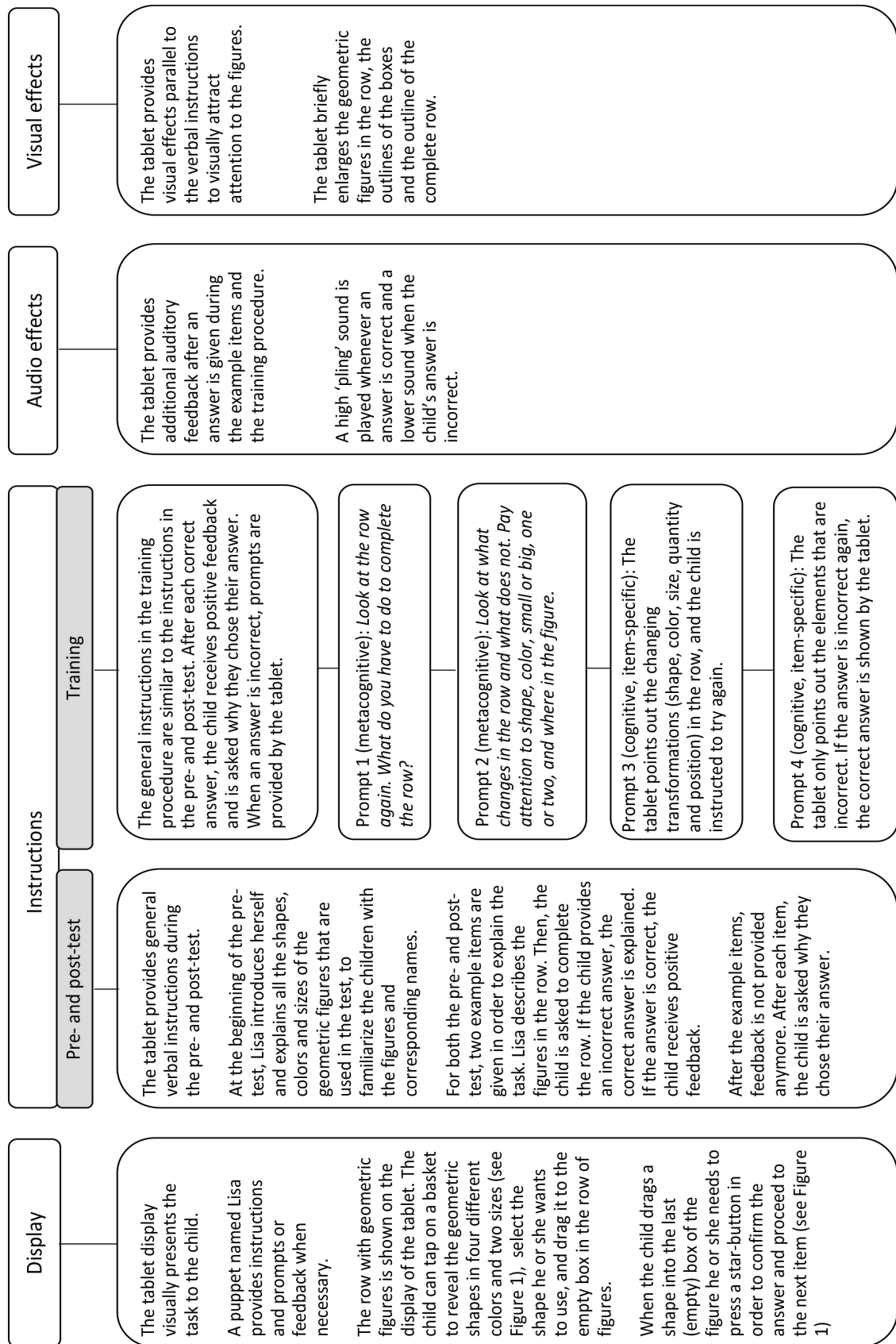


Figure 3. Detailed overview of the computerized series completion

made by the children. A Spearman's rank correlation analysis between the theoretical item difficulty, based on the number of transformations and periodicity, and the measured p -value of the pre-test, revealed an intercorrelation of $r_s = -.81$. This outcome suggests a strong relationship between the theoretical model and the empirically found item difficulty.

Cronbach's alpha was used as a measure for the internal consistency of the set of test items. A Cronbach's alpha of .64 was found for the pre-test. The internal consistency of the post-test was examined separately for the two conditions, training and control. For children in the control condition an alpha of .63 was found, whereas Cronbach's alpha for children in the training condition was .65.

As expected, test-retest reliability analysis revealed that the correlation of the pre-test and post-test scores for the children in the control group was considerably higher ($r = .742, p < .001$) than for the children in the training group ($r = .348, p = .002$).

An aspect of the convergent validity of the series completion test was examined by correlating children's pre-test scores on the series completion test with the Raven's Progressive Matrices scores. Both test scores are considered to measure aspects of inductive reasoning ability. A correlation of $r = .53$ ($p < .001$) between these two measures was found.

Group differences

Prior to examining our research questions, we analyzed whether children in the two conditions did not significantly differ in level of inductive reasoning, using pre-test scores, and age at the start of testing. Two ANOVAs were conducted. Results showed that children in the two conditions did not significantly differ with regard to their initial level of inductive reasoning ($F(1, 162) = .142, p = .707$), nor in age ($F(1, 162) = 2.245, p = .136$). Table 2 provides an overview of the basic statistics of the two experimental groups.

Effect of the dynamic series completion test

We expected that trained children would show more progression in task solving on the series completion test than children in the control group. First, we examined children's progression in accuracy, and conducted a one within-subjects (Session) and one between-subjects (Condition) repeated measures ANOVA, with the number of accurately solved items as the dependent variable. A significant main effect of Session showed that both groups of children, on average, progressed in accuracy from pre-test to post-test (Wilks' $\lambda = .66, F(1, 162) = 83.14, p < .001, \eta_p^2 = .34$). More importantly, children in the training and control condition differed in the level of this progression, as indicated by a significant Session x Condition effect (Wilks' $\lambda = .76, F(1, 162) = 51.19, p < .001, \eta_p^2 = .24$). Consistent with our hypothesis, the trained children showed steeper progression lines from pre-test to post-test compared to children receiving no training, which has been depicted in Figure 4a.

Secondly, we investigated whether the trained children would apply more transformations correctly after training than the control group. A one within-subjects (Session) and one between-

Table 2. Basic statistics of the children in the two conditions (Control and Training).

			<i>N</i>	<i>M</i>	<i>SD</i>
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
Accuracy	Control	Pre-test	84	7.50	2.61
		Post-test	84	7.87	2.36
	Training	Pre-test	80	7.65	2.49
		Post-test	80	10.70	2.59
Transformations	Control	Pre-test	84	80.29	7.52
		Post-test	84	80.54	7.49
	Training	Pre-test	80	81.34	6.61
		Post-test	80	86.96	6.58

subjects (Condition) repeated measures ANOVA, with the total number of correctly applied number of transformations as the dependent variable, revealed a significant Session effect (Wilks' $\lambda = .84$, $F(1, 162) = 31.13$, $p < .001$, $\eta_p^2 = .16$) and, again, as expected, a Session x Condition (Wilks' $\lambda = .86$, $F(1, 162) = 26.06$, $p < .001$, $\eta_p^2 = .14$) effect was found. Children in both conditions showed, as expected, progression from pre to post-test with regard to their correctly applied number of transformations, and children receiving training showed larger progressions in this application of the correct transformations when solving the items (see Figure 4b). These outcomes indicated that the training with graduated prompts was effective in increasing the likelihood that children could solve items

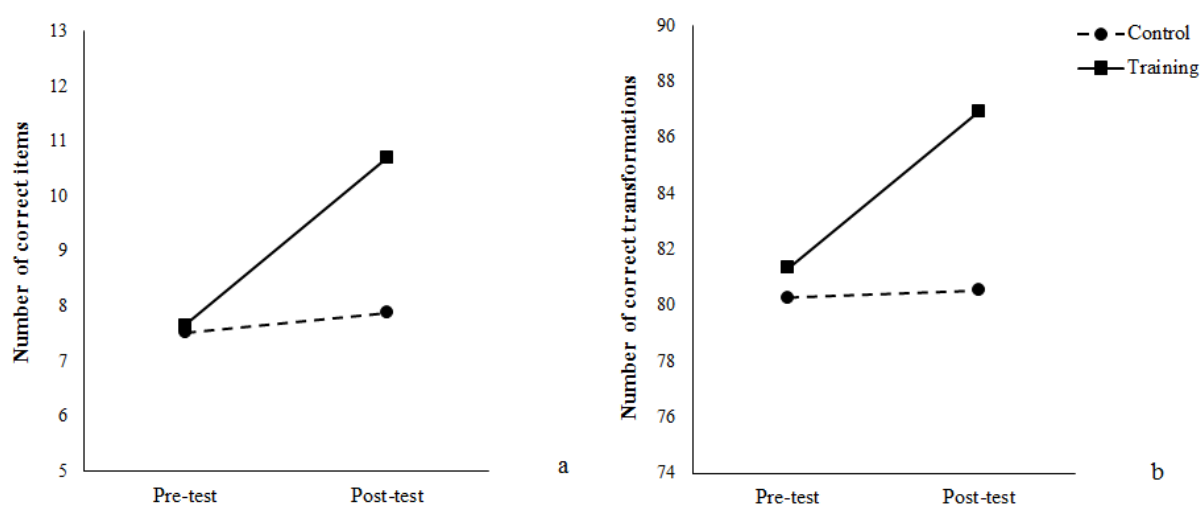


Figure 4. Mean scores of (a) the number of accurately solved items and (b) the number of correctly applied transformations during the pre- and post-test, by condition.

accurately.

Prediction of school achievements

Our next research question focused on the predictive value of our dynamic series completion test in relation to the performance of children on several standardized school achievement tests (Cito). Three ordinal regression analyses were performed, with children's pre and post-test scores, and the number of required prompts as predictors, for the training condition only. Children's scores on Cito Math, DMT and Spelling were used, respectively, as dependent variables. Table 3 depicts the outcomes of these analyses.

Results of the analysis with the mathematic scores revealed that the final model gave a significant improvement over the intercept-only model ($\chi^2(3) = 13.77, p = .003$). In addition, the Pearson goodness-of-fit index also suggested that our final model was a good fit ($p = .376$). The parameter estimates showed a significant relationship between the number of prompts (dynamic measure) and mathematic scores ($\chi^2(1) = 5.97, p = .015$). The coefficient was negative ($-.093$), indicating that fewer prompts needed during training was related to an increase in the odds of obtaining a higher mathematic score. The pre-test and post-test scores showed no significant relation with mathematic scores. Similar results were obtained for the technical reading scores, measured with the Cito DMT test. The final model gave a significant improvement over the intercept-only model ($\chi^2(3) = 8.81, p = .032$). The Pearson goodness-of-fit index also suggested that our final model was a good fit ($p = .821$). Inspection of the parameter estimates revealed a relationship between the number of prompts and technical reading scores ($\chi^2(1) = 3.91, p = .048$). Again, the negative coefficient ($-.077$) indicated that fewer prompts needed during the training was related to increased odds of obtaining a higher technical reading score. Results obtained for the spelling scores showed that the final model was a significant improvement over the intercept-only model ($\chi^2(3) = 10.25, p = .017$). The Pearson goodness-of-fit statistic gave evidence of a good fit of the final model ($p = .835$). The parameter estimates, however, showed no significant relationships between the static and dynamic series completion scores and spelling scores.

Overall, these results suggested that the number of prompts, which is a dynamic measure of our series completion test (e.g., Resing, 2013), showed the highest predictive value for children's performance on both mathematic and technical reading scores. No separate significant predictor for the spelling scores was found, indicating that spelling, as measured with the Cito test, refers to different cognitive abilities than inductive reasoning, as measured with the series completion test.

Teacher ratings

Lastly, the relationship between the teacher ratings (impression of overall school performance and potential for learning) and the results of the series completion test and actual school performance were analyzed.

Table 3. Ordinal regression analyses, with pre-test scores, post-test scores, and number of prompts needed during training as predictor variables, and mathematics ($N = 79$), technical reading ($N = 77$) and spelling ($N = 77$) achievement as dependent variables.

		Estimate	Std. Error	Wald	Sig.
Mathematics	Pre-test	.118	.106	1.223	.269
	Post-test	-.093	.090	1.061	.303
	Prompts	-.093	.038	5.974	.015*
Technical reading (DMT)	Pre-test	.093	.109	.730	.393
	Post-test	-.145	.093	2.433	.119
	Prompts	-.077	.039	3.915	.048*
Spelling	Pre-test	.148	.108	1.885	.170
	Post-test	.053	.091	.335	.563
	Prompts	-.046	.038	1.493	.222

* Significant value ($p < .05$).

First, the data of the *control group* children were examined. Spearman correlations showed that the teacher ratings regarding children's overall school performances and potential for learning correlated significantly, between $r = .308$ and $r = .704$, with the static measures of the series completion test and children's actual school performance as measured with the Cito tasks (see Table 4). For the *trained group*, we found comparable correlation patterns between teacher ratings and children's actual school performance measures (between $r = .582$ and $r = .781$). Teacher ratings also correlated with children's scores on the series completion test. The correlations between, on the one hand, teacher ratings, and, on the other hand, the pre-test (static measure) and the number of prompts (dynamic measure) were significant (between $r = .299$ and $r = -.345$). Non-significant correlations were found between teacher ratings and the post-test of the trained group (dynamic measures) ($r = .142$ and $r = .192$). This unexpected, but interesting, result will be elaborated on in the Discussion section.

2.4 Discussion

The current study sought to investigate the effects of a computerized series completion test using a dynamic testing approach. We compared the accuracy in solving series completion items of children who were trained on a visual-spatial figure series completion test with those of children who did not receive training. In accordance with outcomes reported in previous studies utilizing dynamic series completion tests (e.g., Ferrara et al., 1986; Resing & Elliott, 2011; Resing, Touw, Veerbeek, & Elliott, 2017; Stad, Wiedl, & Resing, 2016), we conclude that the training providing a range of prompts becoming gradually more specific is effective in increasing the likelihood that children can learn to

Table 4. Spearman's rank correlation of the teacher ratings with children's performance on the series completion test and Cito scores.

Teacher ratings	Condition	Pre-test	Post-test	Prompts	Cito math	Cito DMT	Cito spelling
School performance	Control N = 71	.450**	.350*	-	.704**	.389**	.535**
	Training N = 69	.299*	.192	-.264*	.745**	.582**	.689**
Learning potential	Control N = 71	.319*	.308*	-	.675**	.499**	.520**
	Training N = 69	.315*	.142	-.345*	.781**	.588**	.723**

Note. * $p < .05$, ** $p < .001$

solve series completion problems accurately. Compared to the children in the control condition, trained children also showed greater progress in correctly applied transformations, which can be seen as a progression in task-solving strategies by noticing the underlying solving rules (Resing, Bakker, Pronk, & Elliott, 2016).

As opposed to previous studies in the field of dynamic testing, in the current study an innovative computerized dynamic test was used: instructions and feedback were provided by the tablet, instead of the examiner. For children of the age group examined, the tablet and the computerized instructions regarding the dynamic series completion tests, were not difficult to handle. Observation suggested that both the children and examiners enjoyed the way of testing by the tablet: It appeared to be easily administered, and children were excited and seemed motivated to work on the tablet. These observations correlate with earlier research, in which both teachers and children enjoyed working with mobile devices such as tablets (Ciampa, 2013; Furio et al., 2015). Although children seemed motivated to work on the tablet, the examiners observed that, for some children, the pre and post-test, both consisting of 18 items, were relatively long. Perhaps the relatively high difficulty level of the different items, to avoid ceiling effects, contributed to this. The effect of tablet use on children's motivation was not a focus of the current study, but is certainly interesting and worthwhile examining further in future studies comparing paper-and-pencil with computerized dynamic tests.

Furthermore, in line with earlier research findings (Caffrey et al., 2008; Stevenson et al., 2013) one of the dynamic measures of the series completion test, i.e. the number of prompts children needed, predicted the performance of children on the standardized school tests of mathematics and technical reading (DMT) quite well. Seriation tasks are widely used across psychological and educational settings, and are assumed by some researchers to be related to academic skills, such as number comprehension and mathematical concepts (De Koning, Sijtsma, & Hamers, 2003; Kingma,

1981). In part, the lack of a significant relationship between the static pre-test scores on our series completion task with scholastic achievement scores could be explained through the fact that series completion tasks require a form of inductive reasoning, a domain-general skill which is closely related to, for example, general cognitive ability (Goswami, 2012; Klauer & Phye, 2008), whereas the scholastic achievement tests are developed to measure past learning experiences in a specific academic domain.

In general, teacher's estimations of children's potential for learning and their scholastic achievements were strongly related to children's actual school performances as measured with the Cito tasks, which was in line with previous studies (e.g., Bosma & Resing, 2008; Feinberg & Shapiro, 2009; Südkamp et al., 2012). Teachers seem to base their ratings of children's school achievement and potential for learning on the children's scholastic achievement scores. These estimations are, however, only moderately related to the outcomes of the series completion test. Especially weak relations were found with the post-test outcomes of the trained children, which can be seen as a dynamic measure. Although at first sight this seems odd, these low associations could, at least partly, be explained by the fact that dynamic testing can be seen as a way to provide more information about the learning progress of individuals, and their cognitive potential, rather than an overview of what they have learned so far (Elliott, Grigorenko, & Resing, 2010). It may be that teachers base their rating of pupils for a large part on observations and static school test results, which, in turn, predominantly provide an overview of what children have learned so far, which has also been found in previous studies (e.g., Resing, Bosma, & Stevenson, 2012). By solely relying on static test outcomes, they might not acquire enough information about whether and how much a child could learn through training or feedback in a new domain, and do not renew this 'picture of the learning possibilities of the child', but rather build this picture on how much their pupil has learned in the scholastic achievement domain in the past. Past learning experience are, however, not always indicative of potential for learning (Resing, Elliott et al., 2012), see, for example, the Pygmalion in the classroom effect (Rosenthal & Jacobson, 1968). Dynamic testing outcomes, on the other hand, provide information regarding the number and type of prompts children use, their flexibility in using these prompts, their progress after training, et cetera, in a new domain (e.g., Elliott, 2003), and, thereby, can be a source of useful, hands-on information for devising didactic strategies and interventions (e.g., Jeltova et al., 2007). The benefits of dynamic testing for teachers, therefore, lie in the fact that dynamic testing results are not only useful in obtaining information about pupils' educational needs and potential, but also in improving the manner in which these aspects are assessed (e.g., Bosma & Resing, 2008, 2012). Educational or school psychologists, of course, play a large role in administering these tests and, subsequently, explaining the test outcomes to the teachers. In turn, these test outcomes could influence teachers' views on individual children's capabilities, and the way these children are addressed in school. Future studies could examine to what

extent teachers' judgments of children's potential for learning might change as a consequence of having been provided with dynamic testing outcomes.

There are of course also other factors which can influence teachers' judgments, such as behavior of the child that is being judged. For example, teachers have been shown to have more difficulty with rating children whose performance is weaker (Feinberg & Shapiro, 2009). Moreover, teachers might have more difficulty with judging pupils' cognitive skills, such as intelligence or potential for learning, than academic skills (Machts, Kaiser, Schmidt, & Möller, 2016). Investigating the factors that might influence teacher judgment will certainly be a focus of future research.

The current study contributes to the growing field of computerized testing by developing a computerized dynamic series completion test. We carefully constructed our items based on the model of Simon and Kotovsky (1963), and although the data revealed a strong association between the expected theoretical and the empirically found item difficulty, the reliability (in terms of internal consistency) of the test was modest, which might have influenced our research outcomes negatively. Inspection of the difficulty levels of the different items led us to conclude that these items were either relatively easy or very difficult to solve. The lack of items of moderate difficulty could possibly be explained by the fact that for the more difficult items, children have to make a switch in their solving strategies. Children have to focus on more information in the items and find more than one solving rule, and this is what the training teaches them. This possible switch in strategy use may appeal to children's cognitive flexibility skills. Children's cognitive flexibility is shown to play a significant role in increasing children's performance on inductive reasoning tasks (Stad et al., 2016). Children's strategy use and cognitive flexibility, and the supporting role dynamic testing can fulfill in this, will, therefore, be a focus of future research into computerized inductive reasoning assessment using a dynamic testing approach. Moreover, such studies could also investigate the benefits of adaptive dynamic testing, enabling tailoring of items of different difficulty levels to the individual child's zone of proximal development, utilizing the series completion items constructed for the current study to be able to cater to the needs of larger age group.

Another point to take into account when conducting future research into the relationship between scholastic achievement and dynamic testing concerns the manner in which scholastic achievement was measured in the current study. The Cito test outcomes of the children were used as a measure of their scholastic achievement. Although the Cito tests are widely used in the Netherlands as a robust measure of school achievement (Hollenberg et al., 2017), children's test results are categorized into one of five categories. This system may have taken away variance in the test scores, which might have influenced the strength of the correlations found. Therefore, in future studies different (dynamic) tests measuring scholastic achievement need to be used (e.g., Jeltova et al., 2007; Yang, Fu, Hwang, & Yang, 2017).

In conclusion, our newly developed computerized series completion test combines the use of electronic technology and dynamic testing to overcome the limitations of conventional static testing, and as a result, adapts more closely to children's individual needs. In the seamless-learning-setting of the dynamic test of series completion, prompts and scaffolds were provided by the tablet, instead of a human examiner, which is what made the current study innovative. By using a tablet, we have created a modern and cost-efficient assessment tool, which can easily be administered by educational and school psychologists to provide insights into children's individual instructional needs and potential for learning. Dynamic testing utilizing a tablet worked very well in the participating schools. The combination of dynamic testing, which yields more insight into children's potential to learn, and the use of technology can help move education towards a more integrated and effective student-centered learning environment, in which catering to individual children's needs has become the standard, rather than the exception. The advantages of a computerized tablet-administered dynamic test further lie in the possibilities for adaptive testing, for example by means of providing prompts and scaffolds according to the individual needs of testees. Of course, more research is needed in computerized dynamic testing, for example in different domains than inductive reasoning. By linking computerized dynamic tests to the content of the curriculum for specific school subjects, the technology and insights from the current study can be extended and used across different countries and cultures.

Appendix A

Item	Changing transformations and periodicity	Number of elements	Expected theoretical difficulty level	<i>p</i> -value	
				Pre-test → Post-test	
				Control	Training
1	C ⁴	1	1	0.95 → 0.94	0.98 → 1.00
2	G ³ , P ³	1	2	0.83 → 0.95	0.84 → 0.93
3	G ³ , Q ²	1	5	0.65 → 0.81	0.61 → 0.94
4	C ⁴ , S ²	2	3	0.81 → 0.74	0.84 → 0.90
5	S ² , Q ² , P ²	1	4	0.75 → 0.85	0.86 → 0.91
6	C ³ , S ² , Q ²	1	6	0.79 → 0.90	0.76 → 0.91
7	G ³ , C ³ , S ² , P ³	1	7	0.71 → 0.71	0.70 → 0.85
8	G ² , C ³ , S ²	1	6	0.12 → 0.11	0.10 → 0.35
9	G ³ , Q ² , P ³	1	6	0.13 → 0.14	0.20 → 0.29
10	C ² , S ² , P ³	1	6	0.63 → 0.60	0.55 → 0.73
11	G ³ , C ⁴ , P ²	1	6	0.30 → 0.27	0.26 → 0.73
12	C ³ , S ² , Q ² , P ³	1	7	0.08 → 0.08	0.11 → 0.29
13	G ³ , C ⁴ , Q ² , P ²	1	7	0.29 → 0.27	0.24 → 0.58
14	G ³ , C ² , S ² , Q ² , P ²	1	8	0.18 → 0.17	0.26 → 0.39
15	G ² , C ³ , S ² , Q ² , P ³	1	8	0.00 → 0.02	0.01 → 0.01
16	G ² , C ⁴ , S ² , Q ² , P ³	1	8	0.13 → 0.19	0.08 → 0.46
17	G ³ , C ⁴ , S ² , Q ² , P ³	1	8	0.05 → 0.04	0.10 → 0.20
18	G ³ , C ⁴	2	6	0.10 → 0.07	0.18 → 0.26

To solve incomplete series, children have to induce the rules by which the patterns in the series change, by detecting differences and similarities. The number of transformations and the periodicity of change are assumed to influence the hierarchy in item difficulty. The series included 1-5 different transformations: changes in geometrical shape (G), color (C), size (S), quantity (Q), and position (P). Further variability has been created through periodicity of change over two², three³ or four figures⁴. Series with only even (^{2 or 4}) or uneven (³) periodicity are expected to be more salient, and therefore easier to solve. Also the number of elements (number of changing series in an item) included in the series will influence difficulty level. The expected theoretical difficulty of the items, partly deducted from the model of Simon and Kotovsky (1963), has been calculated based on these aspects. In the last columns, the empirically measured *p*-values are depicted.

Appendix B

Setup of training procedure as provided by the tablet: verbal, sound, and visual instructions/ prompts

Pre-test and Post-test	Graduated Prompts Training
<u>Display</u>: The tablet display visually presents the task to the child. 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, 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. <u>Verbal instruction</u>: An animated figure provides general verbal instructions. After each item, the children are asked why they chose their answer. <u>Sound</u>: The tablet provides additional auditory feedback after an answer is given during the example items of the pre- and post-test 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. <u>Visuals</u>: 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.	
	<u>Prompts</u>: 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.
	Prompt 1 (metacognitive): <i>Look at the row again. What do you have to do to complete the row?</i>
	Prompt 2 (metacognitive): <i>Look at what changes in the row and what does not. Pay attention to shape, colour, small or big, one or two, and where in the figure.</i> Prompt 2 is also provided visually:
	
	Prompt 3 (cognitive, item-specific): The tablet points out the changing transformations (shape, colour, size, quantity and position) in the row, and the child is asked to try again.
	Prompt 4 (cognitive, item-specific): The tablet only points out the elements that are incorrect. If the child's answer is incorrect again, the correct answer is shown by the tablet.