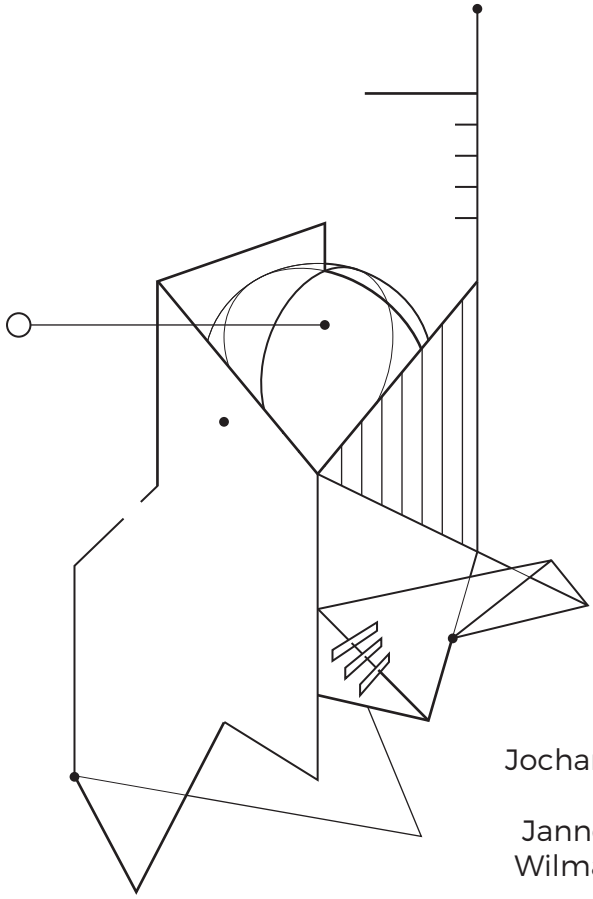
Hidden treasures: Uncovering task solving processes in dynamic testing

Veerbeek, J.

CHAPTER 4

Process assessment in dynamic testing using electronic tangibles



Jochanan Veerbeek
Bart Vogelaar
Janneke Verhaegh
Wilma C. M. Resing

This chapter was published as: Veerbeek, J., Vogelaar, B., Verhaegh, J., & Resing, W. C. M. (2019). Process assessment in dynamic testing using electronic tangibles. *Journal of Computer Assisted Learning, 35,* 127-142.

Abstract

Task solving processes and changes in these processes have long been expected to provide valuable information about children's performance in school. This article used electronic tangibles (concrete materials that can be physically manipulated) and a dynamic testing format (pretest, training, posttest) to investigate children's task solving processes, and changes in these processes as a result of training. We also evaluated the value of process information for the prediction of school results. Participants were N=253 children with a mean age of 7.8 years. Half of them received a graduated prompts training, the other half received repeated practice only. Three process measures were used; grouping behavior, verbalized strategies, and completion time. Different measures showed different effects of training, with verbalized strategies showing the largest difference on the posttest between trained and untrained children. Although process measures were related to performance on our dynamic task, and to math and reading performance in school, the amount of help provided during training provided the most predictive value to school results. We concluded that children's task solving processes provide valuable information, but the interpretation requires more research.

4.1 Introduction

In both clinical and educational settings, cognitive ability tests are often used when questions regarding the overall cognitive or learning abilities of pupils have to be answered (Fiorello et al., 2007). Although these instruments are said to offer the best available prediction of school achievements and to a lesser extent, job performance (Richardson & Norgate, 2015), intelligence test scores are only modestly related to school achievement and, therefore, a great deal of variance in school performance remains unexplained (Fiorello et al., 2007; Neisser et al., 1996; Richardson & Norgate, 2015; Sternberg, 1997).

Intelligence tests have been subject to criticism, because these instruments usually have a static test format, with only one measurement moment, without providing feedback, and are therefore said to measure what a child already knows. In addition, scores on these tests provide only limited information on how children solve the test problems (Campione, 1989; Elliott, Grigorenko, & Resing, 2010). Moreover, to evaluate children's ability to learn, not only already acquired knowledge and skills have to be assessed, but also their potential to learn when the opportunity is presented (Grigorenko & Sternberg, 1998; Sternberg & Grigorenko, 2002). These criticisms led to the development of dynamic testing, which involves testing procedures in which a training session is incorporated to assess the child's response to a learning opportunity (e.g., Kozulin, 2011; Lidz, 2014; Resing, 2013; Sternberg & Grigorenko, 2002; Stringer, 2018). To improve the predictive validity of traditional tests, some researchers argued that an additional analysis of the task solving process would provide valuable information regarding cognitive potential (Resing & Elliott, 2011; Resing, Xenidou-Dervou, Steijn, & Elliott, 2012; Sternberg & Grigorenko, 2002). Both the assessment of the child's progression in task solving, including the use of electronic tangibles, and the evaluation of this task solving process were the foci of the process-oriented dynamic testing procedures used in the current study. In the current paper, task solving processes were defined as the task oriented behaviors children employed during inductive reasoning task solving.

Dynamic testing and graduated prompts procedure

Whereas static tests do not include training beyond repeated instruction or, in most cases, do not contain explanations or feedback regarding the correctness of answers, dynamic testing incorporates an instruction moment in the form of feedback, training or scaffolding. Dynamic testing can be utilized to measure progression in task solving, in terms of accuracy scores on the task considered, but also to assess the processes involved in learning how to solve these problems (Elliott, Resing, & Beckmann, 2018; Haywood & Lidz, 2007; Resing & Elliott, 2011; Sternberg & Grigorenko, 2002). Over the years, several different formats have been developed for dynamic testing (Haywood & Lidz, 2007; Lidz, 2014; Sternberg & Grigorenko, 2002). Formats range from relatively unstructured with a great emphasis on the examiners' possibility to provide unique individualized instruction at any point the examiner deems necessary, to completely standardized (e.g., Campione, Brown, Ferrara, Jones, & Steinberg, 1985; Resing, 1998). Dynamic tests have been implemented in a variety of domains including academic subjects and language development (Elliott et al., 2018), with a range of available testing instruments to target the domain of interest (Haywood & Lidz, 2007; Sternberg & Grigorenko, 2002).

In some of the more structured formats, for example a *pretest, training, posttest* design children are provided with graduated prompts as part of the instruction moment (Campione et al., 1985; Fabio, 2005; Ferrara, Brown, & Campione, 1986; Sternberg & Grigorenko, 2002). This procedure provides standardized help, in the form of hints and prompts, which are presented to children if they cannot solve a problem independently. The graduated prompts approach was originally designed to assess individual differences in the amount and type of instruction needed to elicit the solving of tasks, and was further refined to find the degree of help a child needed to complete a task successfully (Campione et al., 1985; Resing, 1993, 2000). Hints are hierarchically ordered, from general, metacognitive prompts, to concrete, cognitive scaffolds. The method of training was found to lead to greater improvement in task success than regular feedback, especially for the children who had low initial scores (Stevenson, Hickendorff, Resing, Heiser, & de Boeck, 2013). More importantly, both the number of prompts and posttest scores were found to be good predictors of future school success as well as an indicator of learning potential (e.g., Caffrey, Fuchs, & Fuchs, 2008).

Inductive reasoning and series completion

In many static and dynamic testing procedures, inductive reasoning tasks are extensively used. The process of inductive reasoning requires one to detect and formulate a general rule within a specific set of elements (Klauer & Phye, 2008). Inductive reasoning ability is considered a core component of children's cognitive and scholastic development (Molnár, Greiff, & Csapó, 2013; Perret, 2015; Resing & Elliott, 2011), and can be measured with a variety of tasks, such as analogies, categorization, and series completion (Perret, 2015; Sternberg, 1985). In the current study schematic picture series completion tasks were used, in which pictorial series had to be completed by inducing and implementing solving rules. Simon and Kotovsky (1963) identified three central components of the inductive reasoning task solving process; (1) the detection of relations/transformations in the material, (2) the identification of periodicity, and (3) the completion of the pattern.

Series completion tasks can be constructed with a range of contents such as letters, numbers, and pictures. Letters and numbers have a fixed, often familiar relationship to each other. Pictures and colors on the other hand, do not, and, therefore, require more analysis of the sequence to determine the relationship(s), and, in doing so, solve the tasks (Resing & Elliott, 2011). Schematic pictures, as used in the current study, can consist of several combined sets of transformations, which are not necessarily related (e.g., Sternberg & Gardner, 1983), and have a constructed response format. As opposed to multiple choice items, constructed response items were found to be more difficult to solve, but also to elicit more advanced and overt task solving processes on a dynamic test of analogical reasoning in 5- and 6-year old children (Stevenson, Heiser, & Resing, 2016).

Process-oriented testing

When children or adults are first presented with a problem to solve, they, in principle, attempt to understand it by creating an initial problem representation. According to Robertson (2001), the efficiency and accuracy of the task solving process is determined by the quality of this representation. As argued by many researchers, this initial representation is a crucial aspect of performance (Hunt, 1980; Pretz, Naples, & Sternberg, 2003). As problem representation is said to determine the strategies that are chosen to try and solve a problem, an incorrect representation may result in the use of inaccurate

strategies (Alibali, Phillips, & Fischer, 2009; Pretz et al., 2003). The problem representation of a solver can potentially be improved as the result of learning to use new solving strategies. Often, the extent to which improvement is successful is believed to be dependent on the availability and organization of the requested knowledge (Pretz et al., 2003).

Moreover, the notion of "problem space" was introduced by Newell and Simon (1972), as a conceptualization of the problem definition and representation which contain all possible routes to a solution. According to these authors, a problem space can be reduced by restructuring the problem into a set of smaller problems, which is also called "means-ends analysis". This approach is thought to be particularly helpful if no clear solving strategy is available (Robertson, 2001; Weisberg, 2015). The ways in which a solver structures a problem, for example by analyzing the sequence of solving steps or grouping these answering steps in meaningful units, is thought to provide valuable information about individual differences in problem solving. However, most standard cognitive tests have not been constructed to reveal this process information (Richard & Zamani, 2003).

Process-oriented dynamic testing originated from an intention to detect (individual) changes in strategy use as a result of training (Resing & Elliott, 2011), and from the idea that examining strategy use would enable an examiner to assess how a person's solving of a task progresses. Examination of an individual's use of strategies, offering information on which specific strategies might be used more effectively, may provide valuable insight into what a person needs to improve specific task performance (Greiff, Wüstenberg, & Avvisati, 2015). The pivotal role of strategy use in task performance has also been highlighted by Siegler (2004, 2007). He not only found that instability in strategy use over a short period of time is associated with improvement in task performance (Siegler, 2004, 2007), but also that this improvement seems connected to a person's ability to adapt strategy use to the requirements of the situation (Hunt, 1980; Siegler, 1996). He concluded, however, that an individual's global strategy pattern that was displayed throughout learning situations could be characterized by a shift from less to more advanced strategy use (Siegler, 1996; Siegler & Svetina, 2006). Nevertheless, although more expert reasoners appear to use more advanced strategies more frequently, both simple and advanced strategies can produce accurate task outcomes (Klauer & Phye, 2008). Recent studies have stressed that the relationship

between performance and strategy use could be mediated by task difficulty (Goldhammer et al., 2014; Tenison, Fincham, & Anderson, 2014).

In practice, however, process-oriented testing has shown to be challenging, because the sequential solving steps involved can quickly become too much to analyze, or are often difficult to interpret (Zoanetti & Griffin, 2017). With the emergence of computers in the educational and cognitive testing domains, it has become easier to collect data regarding children's process of task solving. Computers allow for monitoring an individual's progress, while providing individual learning experiences (Price, Jewitt, & Crescenzi, 2015; Verhaegh, Fontijn, & Hoonhout, 2007). While the opportunity to analyze problem solving behavior from digital log files has been praised since the early days of computer-based assessment, interpreting these files in a meaningful way has proven to be difficult (Greiff et al., 2015; Zoanetti & Griffin, 2017). As a result, the advantages offered by computerized assessment appear to have hardly been exploited optimally.

Aims and research questions

The current study sought to investigate the possibilities for process-oriented dynamic testing, using various ways of process measurement. By combining these outcomes, we aimed to study the predictive validity of dynamic testing with regard to academic performance. We used a dynamic testing format in which half the participating children were subjected to training between pretest and posttest, to investigate children's potential for learning in both the outcome and the process of solving inductive reasoning tasks. In addition, we tested a rule-based automated scoring method developed to measure changes in problem representation in children's inductive problem solving.

We firstly expected (hypothesis 1) children's problem solving processes and outcomes in series completion to progress to a more sophisticated level. We expected (1a) children to show more accuracy in their series completion solving skills as a result of a graduated prompts training, than as a result of repeated practice (Resing & Elliott, 2011; Resing et al., 2012). Further, we anticipated that (1b) training would lead children to show more grouping activities (separating groups of task elements) to make completion of the series easier, and that (1c) training would lead to more sophisticated verbalized strategy use (Resing et al., 2012). We also expected (1d) a decrease

in the time spent on the task as a result of more familiarity with the type and structure of the tasks as a result of training (Tenison et al., 2014).

Secondly, we investigated children's shifts in the process of solving the series completion tasks as a result of repeated practice and training, by distinguishing subgroups of children based on their initial task solving processes. It was expected that the distribution of children over the subgroups would change from pre- to posttest and that trained children would move towards more sophisticated categories of grouping behavior than non-trained children (hypothesis 2a). We also expected trained children moving towards more advanced verbalized strategy categories than non-trained children (hypothesis 2b).

Thirdly, we expected (hypothesis 3a) process measures to be related to accuracy on the series completion task, and to children's academic performance on mathematics and reading comprehension. The process measures were expected to provide explanatory value for academic performance on mathematics (hypothesis 3b) and on reading comprehension (hypothesis 3c). In line with previous research (Elliott, 2000; Greiff et al., 2013; Zoanetti & Griffin, 2017) we also expected (hypothesis 3d) dynamic test measures (scores) to provide superior prediction over static measures regarding school performance (Caffrey et al., 2008; Resing, 1993).

4.2 Method

Participants

The study employed 253 children, 134 boys and 119 girls (M = 7.8 years; SD = 0.61 years). The children were recruited from twelve second grade classes in nine primary schools, all located in middle class SES regions in the Netherlands. Informed consent was obtained from both the teachers and the parents before testing started. The research was approved by the ethics board of the university. Fifteen children were not able to attend all sessions and therefore their data were not included in the data for analysis.

Design

A pretest posttest control-group design was used (see Table 1 for an overview). A randomized blocking procedure was used to assign children to either the Training (N = 126) or the Control (N = 127) condition. Blocking

in pairs was, per school, based on children's scores on the Raven Standard Progressive Matrices (Raven, Raven, & Court, 1998), collected prior to the pretest session. Per pair, children were randomly assigned to a condition, and, then, were individually tested during four sessions. Children who were assigned to the Training condition received a pretest, two training sessions, and a posttest. Control group children received the same pre- and posttest, but spent an equal amount of time on visual-spatial dot-completion tasks, instead of receiving training sessions. Each session lasted approximately 30 minutes. Sessions took place weekly.

Table 1. Overview of procedures for Training and Control group

	Raven Standard Progressive Matrices	Pretest	Training 1	Training 2	Posttest
Training	X	X	X	X	X
Control	X	X	dots	dots	X

Materials

Raven's Standard Progressive Matrices. To assess the children's level of inductive reasoning ability before testing, Raven's Standard Progressive Matrices was used (Raven et al., 1998). The test consists of 60 items, progressing in difficulty. It requires the children to detect which piece is missing and choose the correct answer out of 6-8 options based on the characteristics and relationships in the item. The Raven test has an internal consistency coefficient of a=.83 and a split-half coefficient of r=.91.

Scholastic achievement. The scores of the Dutch standardized, norm-referenced tests of scholastic achievement [Cito Math (Janssen, Hop, & Wouda, 2015) and Cito Reading Comprehension (Jolink, Tomesen, Hilte, Weekers, & Engelen, 2015)] were provided by the participating schools. These tests have been developed with the purpose of monitoring children's progress on the school subjects. Children's achievement on the test are scored on a scale which ranges from "A" to "E", with "A" scores representing the highest (25%) performance and "D" (15%) and "E" representing the lowest (10%), compared to the average performance of Dutch children of the same age (Janssen et al., 2015; Jolink et al., 2015; Keuning et al., 2015). For two children, a Cito Math score was not available; for 63 children, a Cito Reading

Comprehension score was not provided because their schools did not administer this test. The reliability for Mathematics (M4 [grade 2]), defined in terms of measurement accuracy is MAcc = .93 (Janssen et al., 2015). For Reading Comprehension (M4 [grade 2]), the reliability in terms of measurement accuracy is MAcc = .86 (Jolink et al., 2015).

TagTiles console. A tangible user interface (TUI), TagTiles (Serious Toys, 2011) was utilized for administering the dynamic test. The console consisted of an electronic grid with 12 x 12 fields, which included sensors to detect activity on its surface. The console was equipped with multicolor LEDs, providing visual feedback, and audio playback, used for instructions and prompts during the pre- and posttest and the training.

To use the functionality of computer systems in monitoring behavior and providing automated responses, but not be restricted to the regular computer interface such as a mouse and keyboard, TUIs were developed (Verhaegh, Resing, Jacobs, & Fontijn, 2009). These physical objects allow for natural manipulation, and have electronic sensors built in to use some of the functionality of computers (Ullmer & Ishii, 2000). These TUIs allow for monitoring the task solving process through the physical manipulations of the solver (Verhaegh, Fontijn, et al., 2007). They are easier to use by children, because the physical tangibles do not require any interpretation or representation like PC interfaces do (Verhaegh et al., 2009), thereby allowing for more accurate measurement for assessment purposes (Verhaegh, Fontijn, Aarts, & Resing, 2013; Verhaegh, Fontijn, & Resing, 2013). The console enabled children to work independently (Verhaegh, Hoonhout, & Fontijn, 2007), because it was programmed to provide not only standardized instruction and assistance as a response to the child's actions (Verhaegh, Fontijn, Aarts, Boer, & van de Wouw, 2011), but also to record children's task solving processes step-by-step (Henning, Verhaegh, & Resing, 2010).

Dynamic test of schematic picture series completion. To assess children's task solving process, a dynamic test version of a pictorial (puppets) series completion task was used (Resing & Elliott, 2011; Resing, Touw, Veerbeek, & Elliott, 2017; Resing, Tunteler, & Elliott, 2015; Resing et al., 2012). The puppet task has been designed as a schematic picture series completion task with a constructed response answering format. Each series consists of six puppet figures and the child has to provide the seventh (Figure 1). To solve the task, the child has to detect the changes in the series, by

looking for transformations in the task characteristics and the periodicity of the transformations. From this, the rule(s) underlying these changes have to be induced before the task can be solved (Resing & Elliott, 2011).

The child has to solve each series on the console, using colored blocks with RFID tags. Each puppet consists of seven body pieces, differing in color (yellow, blue, green, pink), pattern (plain, stripes, dots), and head (male, female). The task has varying levels of difficulty, with gradually more changes in the periodicity and number of transformations. The items were presented in a booklet, which displayed one item per page.

Pre- and posttest. The pretest and posttest both consist of 12 items, and are equivalently constructed. Each item on the pretest has a parallel item on the posttest with the same transformations and periodicity (but, for example different colors, patterns, or heads). Both the pretest and the posttest session started with an example item presented and instructed by the console. The two training sessions consisted of 6 items each. Scoring was based on the accuracy of solving the items on the test. The score consisted of the amount of correctly solved items on the test, which could range between 0-12. The overall Pearson correlation between pretest and posttest was (r = .54, p < .001), and was slightly higher for the Control condition (r = .59, p < .001), than for the Training condition (r = .51, p < .001) as would be expected.

Figure 1. Example item of the puppet series completion task

Training. The graduated prompts training procedure that was utilized in the dynamic test includes series that are equivalent to those used on the pre- and posttest. During the two training sessions, the children were given structured and standardized prompts, if they were not able to solve an item independently. These prompts (see Figure 2) were provided by the console, according to a structured, hierarchical procedure that started with general, metacognitive prompts (Resing & Elliott, 2011; Resing et al., 2017, 2012).

The first two prompts were aimed at activating prior knowledge and focusing attention to the task characteristics. If these would not enable the child to solve the series, more specific, cognitive prompts were given, after which, if necessary, a scaffolding procedure was provided, followed by modelling of the solving process. After solving a series, children were asked to tell how they solved the task. The training procedure started with the most difficult items, followed by less difficult items, to enable children to apply their newly learned strategies at the end of the training session (Resing & Elliott, 2011; Resing et al., 2012). To accompany the verbal prompts provided by the console, visual clues were given. The relevant puppet piece would light up to show children where their attention had to be focused, and during the last stage, the verbal modelling was accompanied by colored lights and pre-programmed answering patterns. A human test leader was present to escort the children from and to the classroom. During testing, the test leader recorded the placement of pieces and verbalizations given by the child, providing a backup in case the electronic console would malfunction.

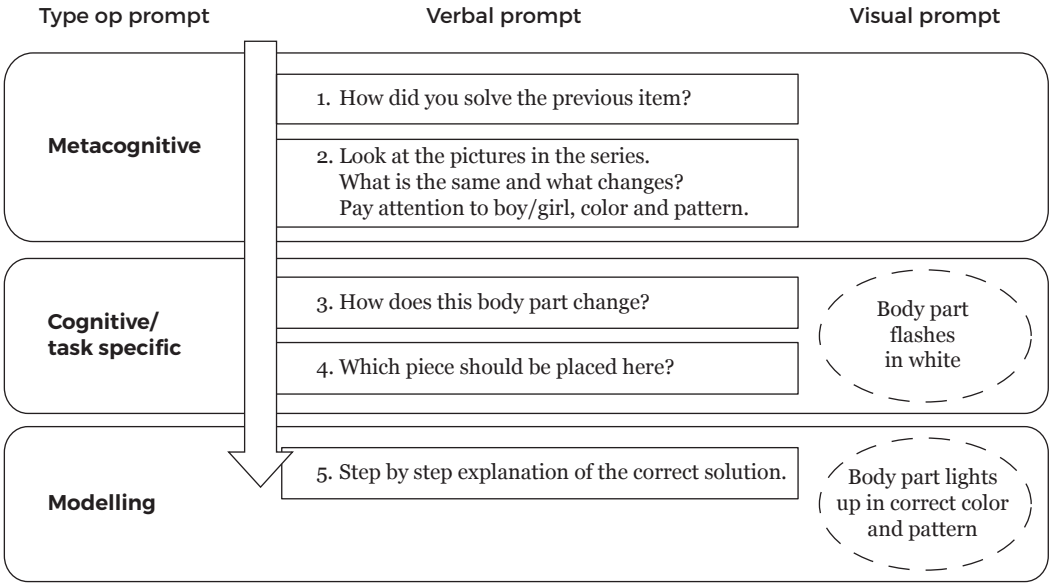


Figure 2. Prompts provided by the electric console during the training procedure of the dynamic test

Scoring

The variables recorded in the log files included the time of placement for each piece, and the identity and placement location of each piece placed on the console surface. In addition, for each item the log files contained the number of correctly placed pieces, completion time, and whether or not the answer that was provided was accurate. The log files were cleared of irrelevant data, such as accidental movement of pieces, or motoric difficulty in the correct placement of the pieces. The relevant data were then imported into SPSS for further analysis. In case of a computer malfunction, data were retrieved from the manually scored hardcopies. Additionally, the manually scored hardcopies included a written record of children's explanations of their solutions. These explanations were also recorded on audio, for which explicit consent was given by the children's parents.

Grouping of answer pieces. The process of solving series problems was operationalized as the way in which the pieces composing the answer were grouped together. Patterns in grouping of answer pieces were assumed to measure whether children were able to divide the problem they had to complete into smaller pieces. In addition, it was analyzed whether these "groupings" were related to the elements and transformations in the series. Which sequences of answer pieces were considered to be adequate for accurately solving the series differed per item, depended on the elements and transformations that were involved in the series. In our study, answer pieces were considered grouped if they were successively placed in an expected sequence. For each item, multiple groups of pieces were discerned that were considered helpful when grouped together. Detailed information on the expected groups can be found in Appendix A. The scoring of the grouping of answer pieces (GAP) was automated in Microsoft Excel, using formulae to identify the sequences of answer pieces per item. For each item, the number of placed groups was divided by the maximum number of groups possible for solving that specific item, which ranged between 2 and 5, depending on the transformations in the item. The final GAP score was composed of the average proportion of groups placed for that testing session.

Additionally, GAP categories were discerned, to make visible shifts in the use of grouping of answer pieces. For each item, the GAP was defined as either full analytical, if all of the expected groups in that item were placed,

partial analytical, if between 50-99% of the expected groups for the item were placed, and non-analytical, if 50% or less of the expected groups for the item were placed.

Children were allocated to a strategy class based on the frequency of GAP scores over all test items. If a single strategy category was used on more than 33% of the items, the child was allocated to the corresponding strategy class. Mixed strategy classes were used if children used two types of GAP in more than 33% of the cases. More information on the categories and classes, and which criteria applied for them can be found in Appendix B.

Verbalized strategies. The children's verbalizations after they solved series items were recorded. These verbalizations were scored according to the three levels used in previous research (Resing et al., 2017). The primary scoring criterion was the extent to which the verbalization included inductive reasoning. If the explanations included none of the transformations necessary to solve the items, and no other explanation that implicitly (e.g. pointing) or explicitly portrayed an understanding of the rules used in the series, the verbalization was appointed to the first group (non-inductive). If transformations or rules were verbalized inductively but incompletely, the verbalization would be categorized in the second group (partial inductive). If a child was able to inductively verbalize all transformations or rules in the task, either implicitly or explicitly, that verbalization would be scored in the third group (full inductive).

Each item's verbalization was scored on its level of inductiveness, and based on these total scores per category, the children were appointed to a strategy class, based on the type of verbalization the children used most or mixed throughout the task. If there was a single type of verbalization used in more than 33% of the items, the child was appointed to the corresponding strategy class. However, if two types of verbalizations were used in more than 33% of the items, the child would be assigned to one of the mixed strategy classes (see Figure 3 for a visual representation, more detailed information can be found in Appendix B).

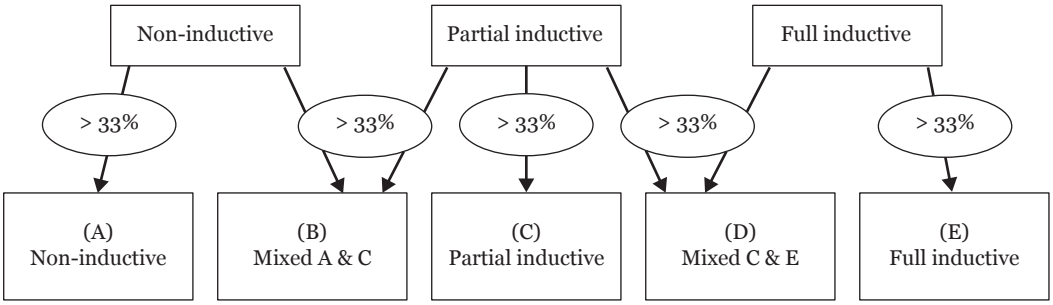


Figure 3. Scoring of verbalized strategy class

Average completion time. To further investigate children's process of solving the series, the item completion times were calculated in milliseconds, based on the time spent between the start of the item, where the console indicated to turn the page of the booklet to the next item, and the end of the item, when children were required to click on the bottom right corner of the console. Out of the completion times, the average completion times were calculated over the full test. For some children (N=18), for which the completion times for one or two items were missing, average time scores were calculated with the remaining items. If the completion times of more than two items were missing, the children (1 at pretest, 3 at posttest) were excluded from the time analyses (N=4).

4.3 Results

Before the hypotheses were tested, preliminary analyses were conducted to check for a priori differences between children in the control and training conditions on Raven scores and age. Univariate ANOVAs, with Raven Standard Progressive Matrices scores and age as the dependent variable and condition (control/training) as the fixed factor, revealed no significant differences in Raven scores (p = .87) or Age (p = .89) between children in both groups. The hypotheses and their corresponding result were provided in Table 8 at the end of the results section for a short overview of our findings.

The effect of training

We expected that children in the dynamic testing group after training would solve the series completion items more accurately than children in the control condition, and would show more advanced patterns in both behavioral and verbal process measures. Means and standard deviations of the dependent variables for the two conditions have been depicted in Table 2 for the pre- and the posttest.

Firstly, a repeated measures ANOVA, with series completion accuracy as the dependent variable, and Condition (training/control) as the between-subjects factor and Session (pretest/posttest) as the within-subjects factor revealed significant main effects for Session and Condition, and a significant interaction effect for Session*Condition (see Table 3 and Figure 4). In line with the expectations, children's series completion solving became more accurate from pretest to posttest, and children who had received training made more progress from pretest to posttest than children who had only been subject to repeated practice.

Secondly, to evaluate the effects of training on children's grouping of answering pieces (GAP), a multivariate repeated measures ANOVA was administered with GAP category (non-analytical, partial analytical, and full analytical) as dependent variable, Session (pretest/posttest) as within-subjects factor, and Condition (training/control) as between subjects-factor. Multivariate effects were found for Session (*Wilk's* $\lambda = .619$, $F(2, 250) = 76.87$, $p < .001$, $\eta^2 = .38$), but not for Condition (*Wilk's* $\lambda = .994$, $F(2, 250) = .791$, $p = .455$, $\eta^2 = .01$), or Session*Condition (*Wilk's* $\lambda = .991$, $F(2, 250) = 1.155$, $p = .317$, $\eta^2 = .01$). Univariate analyses (see Table 3 and Figure 4) per GAP category revealed a significant main effect for Session for non-analytical, partial analytical, and full analytical GAP. These results showed that the use of GAP changed from pretest to posttest. Children used non-analytical GAP less frequently, and partial and full analytical GAP more frequently. However, the graduated prompts training did not result in a faster progression toward more advanced grouping of answer pieces than repeated practice did.

Thirdly, we expected that training would lead to more sophisticated verbalized strategy use. A multivariate repeated measures ANOVA was conducted with Session (pretest/posttest) as within, Condition (dynamic testing/control) as between, factors, and the number of verbal explanations per strategy-category (non-inductive, partial inductive, full

inductive) as dependent variables. Multivariate effects were found for Session (*Wilk's* λ = .799, $F(3, 249) = 20.89$, $p < .001$, $\eta^2 = .20$), Condition (*Wilk's* λ = .965, $F(3, 249) = 2.99$, $p = .031$, $\eta^2 = .04$), and Session*Condition (*Wilk's* λ = .934, $F(3, 249) = 5.83$, $p = .001$, $\eta^2 = .07$). Univariate analyses (see Table 3 and Figure 4) revealed significant main effects for Session for the non-inductive and the full inductive strategy-category, but not for the partial inductive strategy-category. A significant effect for Condition was found for the full inductive strategy-category, but not for the non-inductive and partial inductive strategy-category. Similarly, a significant interaction effect was found for Session*Condition for the full inductive strategy-category, but not for the non-inductive or the partial inductive strategy-category. From pretest to posttest, there was a reduction in the use of non-inductive verbal strategies and an increase in the use of full inductive verbal strategies. More importantly, the trained children showed a sharper increase in the use of full inductive verbal strategies from pretest to posttest than did children in the control condition.

Finally, a repeated measures ANOVA with Session (pretest/posttest) as within-subjects factor, Condition (training/control) as between-subjects factor, and completion time as dependent variable, revealed a significant main effect for Session, but not for Condition, or Session*Condition. Children's completion times became shorter from pretest to posttest, but the training did not lead to a significant difference compared to repeated practice.

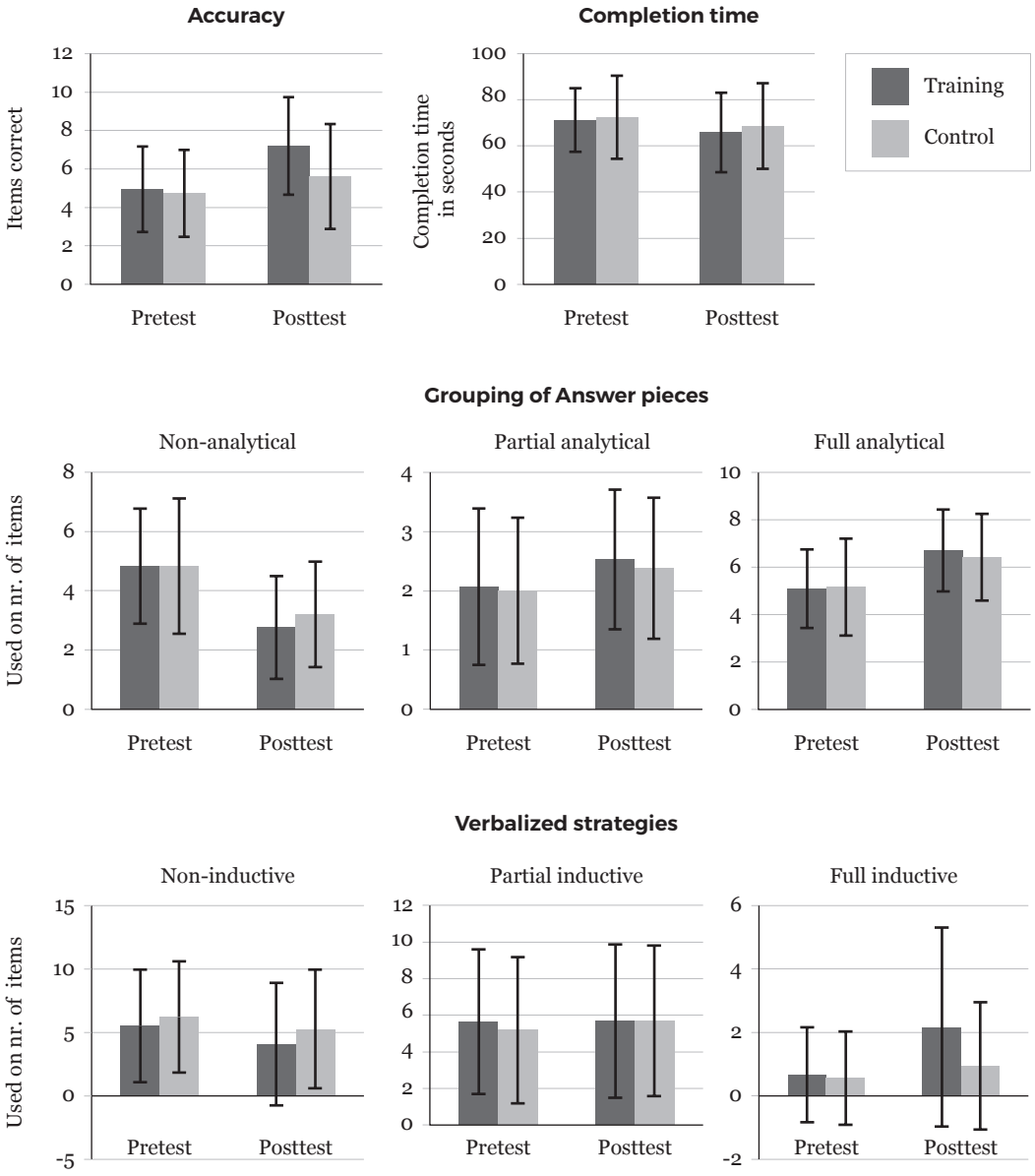


Figure 4. Mean pre- and posttest scores and standards deviations for accuracy, completion time GAP, and verbalized strategies

Table 2. Means and standard deviations for Accuracy, GAP categories, Verbal strategy categories, and Completion Time.

	Trained group (N=126)		Control Group (N=127)	
	Pre *M* (SD)	Post *M* (SD)	Pre *M* (SD)	Post *M* (SD)
Accuracy	4.94 (2.22)	7.20 (2.54)	4.73 (2.26)	5.61 (2.73)
GAP				
Non-analytical	4.83 (1.94)	2.76 (1.74)	4.83 (2.28)	3.20 (1.78)
Partial analytic	2.07 (1.32)	2.53 (1.18)	2.00 (1.23)	2.38 (1.19)
Full analytic	5.10 (1.66)	6.71 (1.73)	5.17 (2.05)	6.42 (1.83)
Verbal strategy				
Non-inductive	5.52 (4.44)	4.07 (4.83)	6.22 (4.38)	5.27 (4.68)
Partial inductive	5.65 (3.96)	5.68 (4.19)	5.18 (4.00)	5.70 (4.12)
Full inductive	.67 (1.50)	2.17 (3.14)	.56 (1.47)	.95 (2.01)
Completion time	71227.85 (13868.59)	65850.17 (17244.00)	72420.26 (18028.97)	68594.56 (18561.95)

Table 3. Results of the Repeated Measures ANOVA's for Accuracy (N=253), GAP categories (N=253), Verbal strategy categories (N=253), and Completion Time (N=249)

	Session			Condition			Session x condition		
	$F(1, 251)$	p	η^2	$F(1, 251)$	p	η^2	$F(1, 251)$	p	η^2
Accuracy	113.10	< .001	.31	11.08	.001	.04	22.15	< .001	.08
GAP									
Non-analytical	153.36	< .001	.38						
Partial analytic	15.30	< .001	.06						
Full analytic	95.91	< .001	.28						
Verbal strategy									
Non-inductive	24.60	< .001	.09	3.30	.071	.01	1.04	.310	.00
Partial inductive	1.35	.247	.01	.248	.619	.00	1.06	.210	.00
Full inductive	51.90	< .001	.17	8.01	.005	.03	17.61	< .001	.07
Completion time	27.26	< .001	.10	.998	.319	.00	.775	.379	.00

Changes in task solving process over time

To further examine the effects of the graduated prompts training procedure on the processes involved in solving series completion, the children were assigned to classes based on their grouping behavior and verbalized strategies used during pretest and posttest. Crosstabs analyses (chi-square tests) were employed to evaluate how children's behavior and verbal solving processes changed over time (Table 4). We analyzed the predicted shifts in GAP by analyzing the relationship between Condition (training/control) and GAP class ((1) non-analytical; (2) mixed 1 & 3; (3) partial analytical; (4) mixed 3 & 5; (5) full analytical). These classes have been described in Appendix B. On the pretest, no significant relationship was found between Condition and the use of GAP ($\chi^{2\ pretest}$ (n=253)= 6.39, p = .172, 40% of the cells have expected count less than 5). On the posttest a significant relationship was found between Condition and the use of GAP ($\chi^{2\ posttest}$ (n=253)= 8.28, p = .041, 25% of the cells have expected count less than 5). As we expected, trained children made more use of more advanced grouping behavior on the posttest than children who had not received training.

Using comparable analyses, we examined the shifts in children's verbal strategy classes ((1) non-inductive; (2) mixed 1 & 3; (3) partial inductive; (4) mixed 3 & 5; (5) full inductive) in relation to the Condition (Training/ Control). The pretest data showed, as expected, no significant effect for condition on the verbalized strategy class ($\chi^{2\ pretest}$ (n=252)= 4.49, p = .344, 40% of the cells have expected count less than 5). However, on the posttest a significant effect for condition was revealed ($\chi^{2\ posttest}$ (n=253)= 14.58, p = .006, 0% of the cells have expected count less than 5). In line with our hypothesis, trained children made more use of more advanced verbal strategies than those who did not receive training.

Table 4. Results for the crosstabs analyses for grouping of pieces and verbalized strategies

		1. Non-analytical	2. Mixed 1 and 3	3. Partial analytical	4. Mixed 3 and 5	5. Full analytical	Missing	Total
Grouping of pieces – Training								
Pretest	Frequency	32	2	40	1	51		126
	Percentage	25.4	1.6	31.7	0.8	40.5		100
Posttest	Frequency	6	0	16	2	102		126
	Percentage	4.8	0.0	12.7	1.6	81.0		100
Grouping of pieces - Control								
Pretest	Frequency	46	1	25	1	54		127
	Percentage	36.2	0.8	19.7	0.8	42.5		100
Posttest	Frequency	18	0	9	3	97		127
	Percentage	14.2	0.0	7.1	2.4	76.4		100

		1. Non-inductive	2. Mixed 1 and 3	3. Partial inductive	4. Mixed 3 and 5	5. Full Inductive	Missing	Total
Verbal explanation – Training								
Pretest	Frequency	54	10	56	4	1	1	126
	Percentage	43.2	8.0	44.8	3.2	0.8		100
Posttest	Frequency	40	7	51	10	18		126
	Percentage	31.7	5.6	40.5	7.9	14.3		100
Verbal explanation - Control								
Pretest	Frequency	57	18	50	1	1		127
	Percentage	44.9	14.2	39.4	0.8	0.8		100
Posttest	Frequency	49	15	51	9	3		127
	Percentage	38.6	11.8	40.2	7.1	2.4		100

Prediction of school achievement test results by static and dynamic test scores

This study also examined the predictive value of process and product measures on the series completion task with regard to school achievement scores on mathematics and reading comprehension. To answer the question whether dynamic measures would provide more predictive value than static (pretest) measures, multiple linear regression analyses were carried out. Math and reading comprehension achievement scores were included as the respective dependent variables and accuracy scores, GAP scores, verbalization class, completion times and number of prompts as predictor variables, for pretest and posttest respectively. Table 5 shows the correlation structure of all variables involved in the various regression analyses.

Table 5. Correlations for process and outcome measures on the puppet task, and Mathematics and Reading comprehension

	Pretest (N=253)			Posttest (N=253)					
				Dynamic testing (n=127)			Control (n=126)		
	Accuracy	Math	Reading	Accuracy	Math	Reading	Accuracy	Math	Reading
Accuracy		.28**	.36**		.37**	.31**		.26**	.31**
GAP	.31**	.20**	.21**	.07	-.06	-.10	.35**	.07	.16
Verbalization	.45**	.11	.22**	.37**	.22*	.15	.41**	.10	.14
Time	.22**	-.03	.02	.30*	.06	-.11	.07	-.11	.07
Prompts				-.72**	-.37**	-.35**			

* $p < .05$. ** $p < .01$.

Hierarchical regression analyses were run on the data of children in the training condition. A first hierarchical regression analysis was conducted with math achievement score as the dependent variable, and the GAP pretest score as the independent variable. This analysis led to a significant model, which explained 4.4 % of variance in Math. In a second model the pretest GAP, verbalization, and completion time were entered as predictors. This model was significant, but did not provide a significant improvement upon the first model. Pretest GAP was the only significant

predictor in this model. A third model in which the pretest accuracy score was added as predictor, led to a significantly better explanation of the variance in math achievement, with an explained variance in math of 9.6%. Accuracy on the pretest of the series completion test and pretest GAP were the only significant predictors in this third model.

A second hierarchical regression was run to analyze the predictive value of the posttest scores regarding the math achievement scores. Model one, with the posttest GAP as predictor, did not show significance. Adding the posttest verbalization and completion time scores as predictors did not lead to a significant model. In a third model posttest accuracy was added as a predictor, which led to a significant model that explained 12.7% of variance in math scores. In this model posttest accuracy was the only significant predictor. An additional model was used, in which the number of prompts provided during training was included as a predictor instead of posttest accuracy. This model significantly explained 12.8 % of the variance in math scores. The number of prompts provided during the training condition was the only significant predictor in this model. In line with our expectations, dynamic (posttest) measures provided more explained variance in math scores (12.7% and 12.8%, respectively) than static (pretest) measures (9.6%).

Similarly, hierarchical regression analyses were conducted regarding the prediction of reading comprehension scores. First, models were tested for the prediction of reading comprehension by the pretest measures. A first model included only pretest GAP score as a predictor, which did not reach significance. In a second model pretest verbalization and completion time scores were added as predictors, which again did not reveal significance. In a third, the pretest accuracy score was added and this model was significant, explaining 12.6 % of the variance in reading comprehension scores. Accuracy was the only significant predictor in this model.

In the hierarchical regression analysis with posttest measures as predictors for reading comprehension, a first model with the posttest GAP score as the only predictor, was not significant. A second model included the posttest verbalization and completion time scores, but again appeared not to be significant. A third model was again tested, with the addition of posttest accuracy as a predictor. This model was significant and explained 12.2% of variance in reading comprehension. In this model, posttest accuracy and completion time were significant predictors. A final model, including number

of prompts provided during training as a predictor instead of accuracy, was significant and explained 14.3 % of the variance in reading comprehension. In this model, again, both number of prompts and completion time were significant predictors to reading comprehension scores. Faster performance on the posttest and fewer prompts provided during the training sessions appeared to be related to better reading comprehension outcomes. It can be concluded that the dynamic testing (posttest) model with number of prompts during training provided marginally more explained variance (14.3%) than did static (pretest) measures (12.6%) to the prediction of reading comprehension. The dynamic model which included accuracy did not provide more explained variance (12.2%).

Table 6. Regression analyses for the prediction of school results for the Dynamic Testing group on the pretest.

Math	Model 1 (F = 6.71*, R^2 = .05)			Model 2 (F = 2.71*, R^2 = .06) $F\Delta$ = .731, $R^2\Delta$ = .01			Model 3 (F = 4.31**, R^2 = .13) $F\Delta$ = 8.58**, $R^2\Delta$ = .06		
(n=125)	*B*	*SE*	*β*	*B*	*SE*	*β*	*B*	*SE*	*β*
Constant	1.67	.83		1.15	.98		1.44	.97	
GAP	3.10	1.20	.23*	3.10	1.20	.23*	2.40	1.19	.18*
Verbalization				.11	.11	.09	-.03	.12	-.03
Completion time				3.93 E-6	.00	.04	-4.52 E-7	.00	-.01
Accuracy							.16	.06	.29**

Reading comprehension	Model 1 (F = 2.53, R^2 = .03)			Model 2 (F = 2.21, R^2 = .07) $F\Delta$ = 2.02, $R^2\Delta$ = .04			Model 3 (F = 4.30**, R^2 = .16) $F\Delta$ = 9.93**, $R^2\Delta$ = .09		
(n=93)	*B*	*SE*	*β*	*B*	*SE*	*β*	*B*	*SE*	*β*
Constant	1.92	1.02		.90	1.29		1.23	1.23	
GAP	2.36	1.49	.16	2.33	1.47	.16	1.44	1.43	.10
Verbalization				.24	.13	.19	.06	.14	.05
Completion time				7.52 E-6	.00	.07	2.47 E-6	.00	.02
Accuracy							.21	.07	.35**

* $p < .05$. ** $p < .01$.

Table 7. Regression analyses for the prediction of school results for the dynamic testing group on the posttest.

Math	Model 1 (F = .397, R^2 = .00)			Model 2 (F = 2.20, R^2 = .05) $F\Delta$ = 3.09*, $R^2\Delta$ = .05			Model 3 (F = 5.46**, R^2 = .16) $F\Delta$ = 14.53**, $R^2\Delta$ = .10			Model 4 (F = 5.53**, R^2 = .16) $F\Delta$ = 14.78**, $R^2\Delta$ = .11		
(n=124)	*B*	*SE*	*β*	*B*	*SE*	*β*	*B*	*SE*	*β*	*B*	*SE*	*β*
Constant	4.45	1.04		3.94	1.06		3.48	1.02		5.74	1.11	
GAP	-.82	1.30	-.06	-1.08	1.30	-.08	-1.23	1.23	-.09	-1.58	1.23	-.11
Verbalization				.20	.08	.21*	.09	.08	.10	.05	.09	.05
Completion time				2.79 E-6	.00	.04	-3.77 E-6	.00	-.05	-2.08 E-6	.00	-.03
Accuracy							.18	.05	.36**			
Nr of prompts										-.05	.01	-.38**

Reading comprehension	Model 1 (F = 1.00, R^2 = .01)			Model 2 (F = 1.68, R^2 = .05) $F\Delta$ = 2.01, $R^2\Delta$ = .04			Model 3 (F = 4.22**, R^2 = .16) $F\Delta$ = 11.27**, $R^2\Delta$ = .11			Model 4 (F = 4.87**, R^2 = .18) $F\Delta$ = 13.72**, $R^2\Delta$ = .13		
(n=94)	*B*	*SE*	*β*	*B*	*SE*	*β*	*B*	*SE*	*β*	*B*	*SE*	*β*
Constant	4.73	1.22		4.92	1.26		4.39	1.21		6.72	1.28	
GAP	-1.56	1.56	-.10	-1.57	1.55	-.11	-1.65	1.47	-.11	-1.95	1.46	-.13
Verbalization				.17	.10	.18	.07	.10	.07	.01	.10	.01
Completion time				-1.02 E-5	.00	-.13	-1.69 E-5	.00	-.22*	-1.52 E-5	.00	-.20*
Accuracy							.18	.05	.36**			
Nr of prompts										-.06	.02	-.41**

* p <.05. ** p <.01. The $F\Delta$ and $R^2\Delta$ of model 4 are based on the change from model 2.

4.4 Discussion

The first aim of the current study was to examine if, and how, dynamic testing, based on graduated prompt techniques and with the use of a TUI, could provide insight into children's potential for learning and their task solving processes. Secondly, our study particularly aimed to investigate the predictive and explanatory value of the process and product measures in a dynamic testing format through rule-based log-file analysis. A new measure for the restructuring of children's problem representations was used, Grouping of Answer Pieces (GAP), along with more often used process measures, being verbalized strategy use (Ericsson & Simon, 1980; Kirk & Ashcraft, 2001; Tenison et al., 2014) and completion time (Dodonova & Dodonov, 2013; Goldhammer et al., 2014; Tenison et al., 2014).

The graduated prompts training, as in previous research with the same dynamic test (e.g. Resing & Elliott, 2011; Resing et al., 2012, 2017) led to more progression in series completion solving performance than repeated practice. The effects of training on the processes children used to solve the tasks revealed a more complex picture. Children's verbalized strategy use became more advanced as a result of training, as evidenced by the increased use of the most advanced, full inductive reasoning strategy-category for the trained children. Improvements were visible in all process measures when children were tested twice, either as a result of repeated practice or training or both. However, children's completion times did not differentially progress under influence of the graduated prompts training. Grouping behavior showed a more complicated picture. The average use of grouping behavior did not appear to progress differently as a result of the graduated prompts training, but the distribution of grouping did show a differential effect after training. It would appear that the graduated prompts training did not affect the level to which the children used grouping behavior, but rather the variability in the use of grouping behavior within the test.

These differential effects for the process measures can be understood in the light of core differences in children's solving processes on the series completion task. On the one hand, verbalizations can be seen as rather task-specific processing, as they are descriptions of the rules underlying the series completion items, representing specific strategies to series completion problem solving. The graduated prompts method most likely provided the

children, if necessary, with detailed task knowledge, which would mean that the more general problem solving structures that are used to solve unfamiliar problems would become less relevant. This notion was supported by the patterns of relations between task success and process measures for the trained children, versus those who had received repeated practice only and children's untrained performance on the pretest. This would be in line with the model proposed by Weisberg (2015), which states that, when solving a problem, the first stage is to search for any available knowledge that could be used for solving the problem. The graduated prompts method procedure provided specific knowledge and methods for solving the series completion task. This knowledge was likely not previously available to the children on the pretest, nor did they acquire it through repeated practice. As a result, untrained performance was dependent on the second and third stages of the model, being domain-general methods, and the restructuring of the problem, respectively (Weisberg, 2015). Grouping behavior, on the other hand, was thought to be a general measure of how children are able to restructure the problem representation, by dividing the task into smaller sub-problems, a form of means-ends analysis (Newell & Simon, 1972; Pretz et al., 2003; Robertson, 2001; Weisberg, 2015). Our data show that most children already used an elementary form of grouping behavior at the pretest, and progressed in doing so when tested twice. This would also explain why GAP, as a measure for restructuring of the problem representation, was no longer related to performance after training. Robertson (2001) distinguished between strong and weak methods of problem solving. Strong methods were described as learned scripts that provide a reasonable certainty of solving the problem correctly. In contrast, weak methods would be methods for the solver to use when no clear method of solving is available. These do not guarantee a correct solution (Newell & Simon, 1972; Robertson, 2001). The graduated prompts training will likely have provided children with strong methods, rendering the use of these weak methods less important to attain a correct solution to the task.

The process measures were weakly to moderately related to accuracy in solving the series completion task. In line with previous expectations voiced in literature (e.g. Elliott, 2000; Greiff et al., 2013; Zoanetti & Griffin, 2017), the process measures used in this study would provide explanatory information on task performance. The rule-based log file analysis was instrumental in uncovering process information, particularly in relation to the restructuring of

the problem representation, by the analysis of the grouping of answer pieces. The predictive value of GAP extended beyond the series completion task performance, to school performance on mathematics and reading comprehension. This supports the notion that process measures, such as GAP, could provide us with more understanding of reasons for not correctly solving the tasks, and subsequently might provide information for intervention (Elliott, 2000; Greiff et al., 2013; Yang, Buckendahl, Juszkiewicz, & Bhola, 2002; Zoanetti & Griffin, 2017). The meaning of the process information, however, seems to differ for each type of process measure. For the grouping behavior, it was found that after training and repeated practice with the task the majority of children progressed toward the most advanced grouping category. This might indicate that low grouping scores could be interpreted as a warning signal. For the verbalizations, on the other hand, even after training, a substantial number of children still provided verbalizations that were classified in the lowest category, because a large group of children were not able to explain how the series should be solved. Only very few children were able to consistently provide complete explanations, and could be identified as the top performers. With regard to completion time, more time spent on the task was associated with better performance. Fast performance would be an indicator that children do not take enough time to acquire information, and control and monitor their actions (Scherer, Greiff, & Hautamäki, 2015).

Previous research has shown superior predictive qualities of dynamic testing for school performance compared to static testing (Caffrey et al., 2008; Elliott et al., 2018), and our findings seem mostly in line with this trend. The dynamic (trained posttest) performance showed a higher predictive relationship for mathematics than did the static (pretest) task performance, as it did in previous research (e.g., Stevenson, Bergwerff, Heiser, & Resing, 2014). For the prediction of reading comprehension, the amount of help provided during training provided more prediction than static test measures, but trained (posttest) performance did not. Furthermore, on the dynamic test, completion time was the only process measure that was related to reading comprehension. Surprisingly, here faster performance was predictive of better reading comprehension scores. This perceived change in relationship between completion time and academic performance may have been the result of a curvilinear relationship, as was found in other domains (e.g. Greiff, Niepel, Scherer, & Martin, 2016), which may have resulted in

a perceived change in relationship when using linear analyses. The other process measures no longer contributed to the prediction of school performance beyond the prediction offered by accuracy. For both math and reading comprehension, the number of prompts children needed during training provided more predictive value than outcome scores.

Of course, this study had some limitations. The use of a constructed response answering format enabled measuring of process indicators, as well as analysis of children's actions through rule-based log file analysis in a manner that would not have been possible in a multiple choice answering format. This poses a limitation to the applicability of the GAP measure, and may prove to be an issue when applying this measure to a more diverse set of tests. We nevertheless would like to encourage future test makers to make use of constructed response answering formats, as it seems to provide useful information, that cannot be obtained from traditional multiple choice tests (Kuo, Chen, Yang, & Mok, 2016; Stevenson et al., 2016; Yang et al., 2002).

It should be taken into account that the current findings were obtained using a series completion task and therefore cannot readily be generalized to any other domains. Similarly, this research was conducted with a single, specific age group, for which inductive reasoning ability is still in full development. Using other age groups in future research could provide us with information on which processes transcend beyond these age limits.

In evaluating the processes involved in solving the series completion tasks, this research used only three separate process measures, which all appeared to measure different aspects of the series completion solving process. Despite using metacognitive prompts during training, this study did not include any measures for level of metacognitive functioning. Future research might identify other factors involved in series completion performance and the training of series completion solving ability. These would not only include cognitive factors such as strategy use and knowledge, but also factors such as metacognitive skills, and emotional and motivational factors. Also, as the task solving process has shown to interact with item characteristics such as item difficulty (Dodonova & Dodonov, 2013; Goldhammer et al., 2014; Tenison et al., 2014), future research should take these item characteristics into account, to gain more detailed insights into the factors that are at play in successfully solving series completion tasks.

Additionally, although this research revealed some indications that process measurement can provide information on both reasons for failure and possible interventions, no clear framework yet exists to interpret these process measures, or connect them to practical and evidence-based interventions. Future research could provide guidelines regarding process data to inform practitioners on the usability of process measures in assessment and intervention. For example, previous research (e.g. Greiff et al., 2016) found that completion time and complex problem solving showed a curvilinear relationship. Future research could focus on non-linear relationships between process measures and performance to provide more information on their meaning.

In conclusion, this research revealed some information concerning the potential value of process-oriented dynamic testing in predicting school results, and the value of process measures for indicating the underlying causes of success or failure on the dynamic series completion task. Dynamic measures could be utilized to provide increased predictive value for school performance. Through using a constructed response answering format, rule-based log file analysis could successfully be administered to provide measures for the restructuring of the problem representation in children. This measure of children's grouping behavior in solving a series completion task, provided predictive value for both performance on the series completion task itself, as well as mathematics performance in school.

Training was found to result in changes in the processes involved in solving the series completion task. Instead of using domain-general methods of solving the tasks, children appeared to make more use of different, learned scripts after graduated prompts training. The various processes involved in solving series completion tasks played different roles in task success, and were influenced differently by training. These factors should all be taken into account when interpreting children's processes in solving tasks, and may need different interventions to remediate. Indeed, the picture that arises from the different processes involved in solving these problems appears to become more complex as we learn more about them, rendering the possibilities for measurement offered by the use of computer more and more necessary in interpreting these measurements.

APPENDIX A.

Grouping of Answer Pieces, groups per item.

For each item, the pieces that were considered adaptive when grouped together, were discerned. The number of groups per item, and which groups applied to which item, are displayed below.

	Pretest		Posttest	
Item	Nr of Groups	Groups	Nr of Groups	Groups
1	3	1. Arms 2. Legs 3. Body	3	1. Arms 2. Legs 3. Body
2	2	1. Arms + Legs 2. Body	2	1. Arms + Legs 2. Body
3	4	1. Arms + Legs 2. ArmLeft + LegLeft 3. ArmRight + LegRight 4. Body	4	1. Arms 2. Legs 3. Body 4. Arms + Legs
4	2	1. Arms + Legs 2. Body	2	1. Arms + Legs 2. Body
5	4	1. Arms 2. Legs 3. Body 4. Arms + Body	4	1. Arms 2. Legs 3. Body 4. Arms + Body
6	3	1. ArmLeft + LegLeft 2. ArmRight + LegRight 3. Body	3	1. Arms 2. Legs 3. Body
7	5	1. ArmLeft + LegLeft 2. ArmRight + LegRight 3. Body 4. ArmRight + BodyRight + LegRight 5. ArmRight + Body + LegRight	5	1. ArmLeft + LegLeft 2. ArmRight + LegRight 3. Body 4. ArmRight + BodyRight + LegRight 5. ArmRight + Body + LegRight

	Pretest		Posttest	
Item	Nr of Groups	Groups	Nr of Groups	Groups
8	2	1. Arms + Legs 2. Body	2	1. Arms + Legs 2. Body
9	2	1. Arms + Legs 2. Body	2	1. Arms + Legs 2. Body
10	4	1. Arms 2. Legs 3. Body 4. Arms + Body	4	1. Arms 2. Legs 3. Body 4. Arms + Body
11	3	1. ArmLeft + LegLeft 2. ArmRight + LegRight 3. Body	4	1. Arms 2. Legs 3. Body 4. BodyLeft + BodyRight + Legs
12	5	1. Arms + Legs 2. Body 3. ArmRight + BodyRight + LegRight 4. BodyLeft + BodyMiddle 5. Arms + BodyRight + Legs	5	1. Arms + Legs 2. Body 3. ArmRight + BodyRight + LegRight 4. BodyLeft + BodyMiddle 5. Arms + BodyRight + Legs

APPENDIX B.

Categories of grouping behavior and verbal strategies.

Scoring of the different categories per item for grouping behavior and verbal strategies, and assignment to classes based on the use of these strategies during the test session.

Grouping behavior	Description of category per item
Full analytical	Based on a GAP score of >99% for an item, which indicates adaptive grouping of the puppet parts, based on the transformations in the item (pieces that go through similar transformations are grouped together) and similarity in other characteristics such as color, pattern, or anatomy (arms, legs, body).
Partial analytical	Based on a GAP score of 51-99% for an item, which indicates some use of adaptive grouping, but not yet consistently using all of the transformations and characteristics of the item to structure the solving process.
Non-analytical	Based on a GAP score of 50% or lower, as an indicator of idiosyncratic solving which is not based on the analysis of the item characteristics, but instead an unplanned or inflexible approach to solving the task.
Verbal strategy	**Description of category per item**
Full inductive	An inductive description of all the transformations in the task is provided, which could be completely verbal, or partially verbal with support of implicit explanation components such as pointing.
Partial inductive	The child is able to provide some inductive explanation of the transformations in the series, but does not explain all transformations that are necessary to successfully complete the task.
Non-inductive	No inductive explanation is provided, but instead the explanation is either lacking (“I don’t know”), or based on information other than the relevant item characteristics (“I like pink”).

Based on the most frequently used categories of grouping behavior and verbal strategies, children were allocated to classes which reflected their most frequently used style of solving the items.

Grouping class	Verbalization class	Rules for classification
1. Non-analytical	1. Non-inductive	Non-analytical/non-inductive behavior was used the most and at least in >33% of the items on the testing session (pretest/posttest)
2. Mixed 1 & 3	2. Mixed 1 & 3	Both non-analytical/non-inductive and partial analytical/partial inductive strategies were used on more than 33% of the items
3. Partial analytical	3. Partial inductive	Partial analytical/partial inductive behavior was used the most and at least in >33% of the items on the testing session. Also included in this class were children that used both non-analytical/non-inductive and full analytical/full inductive strategies in >33% of the items, and children that used all 3 categories equally much
4. Mixed 3 & 5	4. Mixed 3 & 5	Both partial analytical/partial inductive strategies and full analytical/full inductive strategies were used on more than 33% of the items
5. Full analytical	5. Full inductive	Full analytical/full inductive behavior was used the most and at least in >33% of the items on the testing session