The handle [http://hdl.handle.net/1887/71235](http://hdl.handle.net/1887/71235) holds various files of this Leiden University dissertation.

**Author:** Veerbeek, J.

**Title:** Hidden treasures: Uncovering task solving processes in dynamic testing

**Issue Date:** 2019-04-11
CHAPTER 3

Pretest versus no pretest
An investigation into the problem-solving processes in a dynamic testing context

Abstract
Proponents of dynamic testing have advocated its use as a replacement or addition to conventional tests. This research aimed to investigate the effects of using versus not using a pretest on both the outcome on the posttest and the processes used in solving inductive reasoning tasks in dynamic testing using a graduated prompts training. Sixty-seven 7- to 8-year-old children were assigned to either a group that received a pretest or a group that did not receive a pretest, using a randomized blocking procedure. No significant differences were found between both groups of children on posttest accuracy, process measures, number of hints needed during training, amount of time needed for testing, and the prediction of school related measures. This article concluded that the decision of whether or not a pretest is necessary should be based on the research question to be answered because it does not appear to influence posttest results.
3.1 Introduction

Dynamic testing has long been proposed as a replacement of, or an addition to, conventional (static) tests. Unlike conventional tests, dynamic testing uses a learning situation, such as a training or feedback within the testing procedure. Proponents of dynamic testing have argued that this method provides more useful information regarding the learning potential of a child, and how a child can be supported to improve his or her learning, compared to conventional tests (Elliott, Grigorenko, & Resing, 2010). Dynamic testing can be used to investigate why children are not performing closer to their cognitive potential and provide information on what intervention might help to further realize their potential (Haywood & Lidz, 2007; Sternberg & Grigorenko, 2002). Although these advantages appear quite welcome in educational practice, several questions have been voiced over the years. To date, dynamic testing has not become as widely used as some theorists would have hoped or expected (e.g., Hessels-Schlatter & Hessels, 2009; Karpov & Tzuriel, 2009). An issue that has been expressed repeatedly, concerns the possible test-retest effects caused by the use of a pretest, and the threats these pose to the validity and reliability of the test (Kim & Willson, 2010; Klauer, 1993; Sijtsma, 1993). The aim of this research was to investigate the effect of using a pretest within the inductive reasoning domain, compared to not using such a pretest. This effect was studied regarding both the level of quantitative measurement of posttest results and the level of qualitative measurements by investigating the processes involved in answering the posttest items.

Graduated prompts

As a response to the criticized lack of standardization in dynamic assessment, several standardized methods of dynamic testing were developed. One such method is the graduated prompts approach, in which standardized help is provided in a predetermined manner during the training phase. Graduated prompts approaches (Campione & Brown, 1978; Resing, 1993, 2000) are aimed at finding the minimal number of prompts a child needs to be able to solve a task. The graduated prompts approach usually uses a pretest/training/posttest format. During the pre- and posttest, the child is expected to work independently, without the provision of help. During the
training phase, if a child is unable to solve a problem independently, help is offered starting with general, metacognitive prompts. If these do not lead the child to provide the correct answer, more specific, cognitive prompts are provided. For children who are still unable to solve the tasks with the use of these specific, cognitive prompts, eventually a scaffolding procedure is offered, which models the process of solving the tasks to the child (Resing & Elliott, 2011; Resing, Xenidou-Dervou, Steijn, & Elliott, 2012). The graduated prompts method was shown to be an effective training method, leading to a greater increase in the number of correctly solved items than regular feedback (e.g., Campione & Brown, 1987; Ferrara, Brown, & Campione, 1986; Resing, Tunteler, de Jong, & Bosma, 2009; Stevenson, Hickendorff, Resing, Heiser, & de Boeck, 2013). The number of prompts given, together with the posttest score, has shown to be an accurate indicator of the child’s learning potential (Resing & Elliott, 2011).

**Test-retest effect**

In dynamic testing, two frequently used procedures can be distinguished. The first includes feedback or intervention as a reaction to the child’s answer on each item of the test. The second, which is often preferred in research, follows a pretest/training/posttest format. During the pre- and posttest, no feedback or intervention is provided. Between the two test sessions, a training session is provided, which consists of instruction, feedback, or prompts (Sternberg & Grigorenko, 2002). The graduated prompts approach typically uses a pretest/training/posttest design (Klauer, 1993). Criticism on dynamic testing has mainly been focused on the measurement of change between pre- and posttest (Sternberg & Grigorenko, 2002). Kim and Willson (2010), however, concluded in their evaluation of the use of pretesting in educational settings and research that using a stand-alone pretest could lead to problems with, among other things, the internal validity of testing.

An issue that has often been mentioned in research is the difficulty in separating the effects of training from the effects of retesting (Klauer, 1993). Because pretesting can influence scores on all types of constructs, the exact effects of using a pretest will be difficult to predict (Kim & Willson, 2010), and, most probably, will be different across psychological testing domains (Klauer, 1993). In research, this problem has been countered by using a randomized blocked control group design. The use of a pretest may not only
cause higher scores on the posttest but also lead to the measurement of a qualitatively different construct on the posttest. Training may influence the solution process which, as a consequence, may be different from pre- to posttest. This implies that a different construct or ability may be measured on the posttest compared to the pretest (Hessels, Vanderlinden, & Rojas, 2011; Sijtsma, 1993; Wiedl, Schöttke, Green, & Nuechterlein, 2004). Especially for children with intellectual disabilities, there is the possibility that the solution process and strategy use on the posttest are very different from those at pretest, leading to the conclusion that the measured construct is different on both tests (Hessels et al., 2011; Tiekstra, Hessels, & Minnaert, 2009). The effect of pretesting seems to vary among different types of tasks as well. Fluid reasoning, for example, has been claimed to be more strongly influenced by test-retest effects than crystallized intelligence items (Klauer, 1993).

### Problem solving and reasoning

Problem solving is generally seen as an expression of cognitive ability (Richard & Zamani, 2003). Problems have many appearances and are encountered equally in daily life, in school, and in formalized testing. In cognitive ability research, preference is often given to well-defined problems, such as inductive or deductive reasoning tasks, because well-defined problems provide a clear definition of what is being measured (Robertson, 2001). Many cognitive ability or learning tests, both static and dynamic, use some type of inductive reasoning task because inductive reasoning ability is considered to be a good indicator of general cognitive ability (Molnár, Greiff, & Csapó, 2013; Resing & Elliott, 2011). Inductive reasoning tasks can be solved by inferring a rule or rules regarding differences and communalities in the elements of the task and in the relations between these elements (Klauer & Phye, 2008). Inductive reasoning is often seen as a key ability involved in learning and transfer, as learning and transfer both depend on the ability to detect rules and generalize specific knowledge, abilities and skills to other situations and domains (Klauer, Willmes, & Phye, 2002; Perret, 2015; Resing et al., 2012). Over the years, several types of inductive reasoning tasks have been designed and used as a measure of cognitive functioning and intellectual ability. Inductive reasoning is involved in, for example, categorizations, analogies, and series completion. According to Sternberg (1985), each of
these types of tasks require specific types of inductive reasoning ability, thereby asking the problem solver to use specific skills and strategies.

The process of problem solving has been studied and described extensively, mostly building on the pioneering work of Newell and Simon (1972). Before a person can start solving a problem, the problem and all the information surrounding the problem has to be represented internally. This initial problem representation not only determines how a person approaches the problem-solving process, but its quality also affects the efficiency and accuracy of the problem-solving process, making it a crucial aspect of problem-solving performance (Hunt, 1980; Pretz, Naples, & Sternberg, 2003; Robertson, 2001). Therefore, problem representation will guide the selection of strategies for solving the problem. Strategy use and problem representation are said to show an interactive relationship, in which problem representation leads the choice of strategy use, and the availability of strategies influences the problem representation (Alibali, Phillips, & Fischer, 2009).

Sternberg (1985) described problem representation in the domain of inductive reasoning as a metacomponent to problem solving, and defined it as an executive process to structure the problem-solving process. Contrary to strategy use, which is thought to be quite variable across and within different problem-solving situations (Siegler, 1996, 2007), metacomponents are thought to be relatively consistent across tasks (Pretz et al., 2003; Sternberg, 1985). Problem definition and representation were conceptualized by Newell and Simon (1972) as the problem space. According to these authors, the problem space contains all possible outcomes of the solving process and was thought to be reduced by restructuring. Several researchers (e.g., Ericsson, 2003; Newell & Simon, 1972; Robertson, 2001) have indicated that the level of familiarity with the type of task has a profound effect on the way the problem solver approaches the problem. Prior experience with the same or similar tasks may provide the problem solver with methods and strategies to solve the task, either from memory or by transfer (Robertson, 2001; Weisberg, 2015). According to Newell and Simon, if no clear problem-solving strategy is available, heuristics can be used as a guide to restructure the problem space. Heuristics can be described as general but inaccurate approaches that can assist problem solving and which may lead to a quick solution when no specific strategies are available. Dividing the
problem into a set of smaller problems is such a heuristic, also known as means-ends analysis (Robertson, 2001; Weisberg, 2015). Restructuring the problem space was found to be closely related to performance and transfer. Furthermore, the way in which the problem is restructured was found to influence strategy use (Alibali et al., 2009).

Assessing the problem-solving process

Conventional tests are usually focused on the outcome of the problem-solving process, but do not provide information about the processes themselves (Richard & Zamani, 2003). Process-oriented dynamic assessment was designed to enable the measurement of the problem-solving process within a dynamic testing format. Investigating the process of problem solving and the change in this process as a result of training may provide information about how a child learns and what type of intervention a child needs to improve learning and performance (Resing & Elliott, 2011; Resing, Touw, Veerbeek, & Elliott, 2017; Resing et al., 2012).

Despite its possible advantages, process-oriented dynamic assessment has not always been optimally used in practice because the information that results from testing can easily become greater than can possibly be analyzed. The method that has been used most widely in process-oriented research relied on the analysis of the verbal reports, or verbalizations, provided by the problem solver during or immediately after solving the task. Although this method has been criticized in the past (e.g., Feldon, 2010), it is generally seen as providing useful information about the problem-solving process (Tenison, Fincham, & Anderson, 2014). However, advances in measuring the process of problem solving have gone hand in hand with technological advances (Ericsson, 2003). Improved computer technology, more user-friendly hardware solutions, and enhanced software allow for both the monitoring and analysis of test performance (Khandelwal & Mazalek, 2007; Verhaegh, Fontijn, Aarts, & Resing, 2013; Verhaegh, Hoonhout, & Fontijn, 2007). With the emergence of new technology in education, possibilities for large-scale applications of process measurement have come closer to reality (e.g., Siemens, 2013). Automated analysis of the restructuring of the problem space might offer a promising theory-driven measure that could be incorporated within approaches such as learning analytics.
**Aims of the current study**

This study aimed to investigate the effects of using a pretest in a dynamic test within the inductive reasoning domain, both on the product measures on the posttest and on the process measures on the posttest of a dynamic series completion task. To evaluate the process of solving the series completion items, multiple process measures were analyzed, such as the children’s verbalizations about their problem-solving process, time spent for planning and analyzing the problem (Hessels et al., 2011), and the order of solving parts of the problem which is thought to indicate to what level the child uses adaptive restructuring of the problem space. By using a broader range of process measures, we aimed to provide a balanced picture of the effects of using a pretest on the construct validity of the posttest.

First, this article investigated whether the use of a pretest would lead to differences in performance on the posttest. The graduated prompts training used in this research was found to be a very effective training method in previous studies (e.g., Ferrara et al., 1986; Resing & Elliott, 2011; Resing et al., 2009; Stevenson, Heiser, & Resing, 2016) and was expected to compensate for any prior experience from a pretest. It was therefore expected that no significant differences would be found between the children who did and who did not receive a pretest. This was investigated on task accuracy, the use of adaptive restructuring in the order of parts of the problem to be solved, the verbalized strategy, and the percentage of time taken for analyzing the task and planning of the answering process on the posttest of a series completion task.

Second, we expected that the children who did not get a pretest would need the same amount of hints on the training sessions because the pretest was not expected to significantly change the children’s inductive problem-solving proficiency (Hessels et al., 2011; Schorno, 2013). More specifically, it was expected that children who did not receive the pretest would not significantly differ in their need for hints on the first or second training session. Children’s understanding of the type of problems used would not be influenced by whether or not the children received a pretest because retesting only seems to improve superficial familiarity with the type of task (Schorno, 2013).

Third, regarding the time needed to complete the testing cycle, it was expected that the first training would last significantly longer for the children who had not received a pretest than for the children who had received a
pretest, as a result of a lack of familiarization (Schorno, 2013). We expected no significant differences in the time children needed to complete the second training and the posttest between the children who had and who had not received a pretest. We also expected that the full testing cycle would not significantly differ in the total time taken between both groups of children, despite the difference in the number of sessions. Although the children who did not receive a pretest would save time on not having to make the pretest, the additional time needed on the other phases was expected to compensate for the difference in number of sessions.

Fourth, the prediction of posttest accuracy from process measures was expected to not significantly differ for both groups of children. It was also expected that the pattern of correlations with external measures for cognitive functioning (i.e., school performance) would not reveal any significant differences between both groups of children because the graduated prompts training was expected to equalize the factors involved in posttest success and determine the construct validity of the posttest by providing strategies to successfully solve the posttest items (Hessels et al., 2011; Tiekstra et al., 2009).

Fifth, the results were, in an explorative manner, analyzed at item level. Analyzing averaged scores over multiple items could lead to a distorted picture of the importance and diversity of process components such as strategies (Siegler, 1987). It was expected that the process measures would significantly contribute to the prediction of item success. Whether or not a child received a pretest was expected to play no role in the prediction of posttest item accuracy.

3.2 Method

Participants

The participants for this study were 67 children, 28 boys and 39 girls ($M = 7.9$ years, $SD = 0.40$ years). The children were recruited from Grade 2 classes in regular primary schools in The Netherlands. Informed consent was obtained from the parents before testing started. Five children were not able to participate in all the sessions because of illness and were excluded from the sample.
**Design and procedure**

For this research, a modified pretest/posttest control group design was used. Table 1 contains a graphic representation of the design. Children were matched based on their scores on the Raven Standard Progressive Matrices test. Based on this matching procedure, children were randomly assigned to either the Pretest or the No Pretest Condition for the dynamic series completion puppet task. Children in the Pretest Condition received a pretest, two training sessions, and a posttest. Children in the No Pretest Condition did not receive a pretest but did receive the two training sessions and the posttest.

The Raven Standard Progressive Matrices were completed in class, where each child worked from an individual booklet and answering sheet. The series completion puppet task was completed individually by the children in a separate room. The tasks were executed on an electronic console that provided all the necessary instructions, feedback, and training. The examiners were not involved in the testing procedure but were present to record the results on hardcopies to provide a backup in case the console did not properly record the results and to escort the children from and back to class. All sessions took around 30–45 min and were performed with approximately 1 week in between sessions.

<table>
<thead>
<tr>
<th>Table 1. Overview of procedures for Treatment and Control group.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Raven</td>
</tr>
<tr>
<td>-------</td>
</tr>
<tr>
<td>Pretest</td>
</tr>
<tr>
<td>No Pretest</td>
</tr>
</tbody>
</table>

**Materials**

*Raven Standard Progressive Matrices.* The Raven Standard Progressive Matrices were used to obtain a global estimate of the children’s inductive reasoning ability (Raven, Raven, & Court, 1998). The test consists of 60 items which progressively become more complex and are generally considered as a sound test of inductive reasoning ability (Perret, 2015).

*Series Completion Puppet Task.* The main task used in this research was the dynamic series completion puppet task (Resing & Elliott, 2011; Resing et al., 2017). To find the correct answer, the child first had to identify the relations between the elements of the series. Second, the child had to find the
periodicity in the series, to identify what constituted a full cycle of the pattern. Finally, the child had to complete the pattern, by formulating a final rule for all positions that completed the sequence (Resing & Elliott, 2011; Simon & Kotovsky, 1963). The series used in this research consisted of six puppets. Each puppet in turn consisted of eight pieces. The dynamic series completion puppet task was designed as a dynamic test of inductive reasoning. To this end, it used a pretest/training/posttest design. The pre- and posttest contained 12 items and started with an example item. The pre- and posttest items were isomorph; they were designed to have the same level of difficulty and required the same reasoning processes. Within the pre- and posttest, the items were presented in an increasing level of difficulty.

The training phase consisted of two training sessions, which both contained six items. The training sessions used Resing’s (1993, 2000) graduated prompts format. The prompts were provided in a hierarchical order, starting with very global, metacognitive prompts, through cognitive, task-specific prompts. If a child was not able to use these prompts to provide a correct answer, a scaffolding procedure was offered, modeling the process of solving the task step by step. Help would only be offered if and when a child was not able to solve the task independently.

Electronic Tangible Console. The dynamic series completion puppet task was administered with the use of an electronic tangible console, called the TagTiles system (Serious Toys, 2011). This console was used as the surface on which the child was to construct the answer to the series completion task, using tangible blocks that were enhanced with electronic radio frequency identification (RFID) tags. The console had a 12 x 12 electronic grid which detected and recorded the placement and movement of the blocks. Through audio facilities, the console provided the child with all the necessary instructions and prompts. It also provided visual feedback with built-in multicolor LEDs. The use of the TagTiles system enabled full standardization of the instructions and training procedure (Verhaegh et al., 2013; Verhaegh, Fontijn, & Hoonhout, 2007). All recorded data on the console was saved into log files on secure digital cards, which could be used to import the data on a computer for further analysis.

Scholastic Achievement Tests. As a measure of scholastic achievement, the standardized test results of a biannual Dutch scholastic achievement test (Cito) were obtained from the schools. These norm-referenced tests (Keuning
et al., 2015) are widely used in schools throughout The Netherlands. They are administered in January and June of each year, to monitor children’s progress in school subjects (Janssen, Hop, & Wouda, 2015; Jolink, Tomesen, Hilte, Weekers, & Engelen, 2015). For this research, the achievement test scores for Mathematics (Janssen et al., 2015) and Reading Comprehension (Jolink et al., 2015) were used. Children’s scores on both these tests can range from A to E. An “A” score indicates very good performance, as the child’s performance scores within the top 25%, “B” scores indicate good performance in the subject, with a score that falls between the 26th and the 50th percentile, “C” scores are considered as sufficient, as the child scores within the 51st and 75th percentile. Within the lower range, the tests distinguish two categories, with “D” indicating weak scores, which score between the 76th and 90th percentile, and finally “E” scores, which represent very weak scores because these correspond to the lowest 10% scores (Janssen et al., 2015; Jolink et al., 2015; Keuning et al., 2015).

**Scoring and analyses**

After testing, all data were imported into Microsoft Excel, where it was cleared of irrelevant data such as accidental movements of the blocks. All relevant data were entered into SPSS, and, if necessary, missing electronic data were retrieved from hardcopies made by the examiners. All data could be retrieved from the log files or the hardcopies, except for the time intervals used to calculate the planning time and total time. For children where the time intervals of an item were missing (maximally three items per testing session), the planning time for the full test was averaged over the remaining items. Missing total times were treated as missing and excluded from the analyses.

**Accuracy.** The child’s accuracy score in answering the series completion puppet task items was based on the number of correct body parts the child placed during the test items. Each puppet consisted of eight body parts. Because the pre- and posttest both consisted of 12 items, this score could theoretically range between 0 and 96.

**Number of hints.** The total number of hints a child needed per training session was calculated. Each training session consisted of six items, with a maximum of four possible hints per item. Per session, this amounted to a maximum score of 24 hints.
Grouping of the answer pieces. To investigate the restructuring of the problem space, the placement of the answer pieces was analyzed and scored on the use of adaptive grouping of the pieces. Adaptive grouping of the answer pieces (GAP) is considered to indicate restructuring of the problem through the use of a division of the problem into smaller subproblems (Newell & Simon, 1972; Pretz et al., 2003; Robertson, 2001). All items were analyzed to determine which combinations of pieces would be considered helpful when grouped together. The sequence of placement of the puppet pieces would be considered “grouped” if pieces that went through the same transformation or showed similarities in color, pattern, or anatomy were placed in immediate succession of each other. Every puppet piece had its own identification number, so subgroups in sequences could be identified. The numbers ranged from one to eight in the following order: (1) head, (2) left arm, (3) right arm, (4) left body, (5) middle body, (6) right body, (7) left leg, (8) right leg. The number of adaptive groups a child used was divided by the number of possible groups in that item, leading to a score that indicates the proportion of groups used to structure the answering process. The scoring of the grouping of answer pieces was automated by the use of multiple formulae in Microsoft Excel. An example item of the puppet series completion task was displayed in Figure 1.

![Figure 1. Puppet series completion item](image)

For this item, three groups of puppet pieces were distinguished. The first group of puppet pieces consisted of the three body parts because they move through the same transformation and can be grouped together based on the puppet anatomy. The second group of puppet pieces for this item entailed the left arm and left leg, which go through the same transformation. The third was formed by the right arm and right leg, which also
go through the same transformation. The heads were left out of the GAP measure because a single piece cannot be grouped. The number of groups per item ranged between two and five.

**Verbalized strategy use.** Children’s problem-solving strategies were assessed by requiring them to explain how they solved the problem after children had provided an answer. The verbalizations of the children were recorded on audio and scored on their level of inductive explanation of the rule required to solve the task. Three levels of verbalizations were discerned, (a) noninductive, (b) incomplete inductive, and (c) full inductive, as depicted in Figure 2. Based on their verbalizations per item, a score was calculated for the full pre- and posttest. This verbalization class was scored on a 5-point scale and was based on the type of verbalization that was used for more than 33% of the items by the child for that phase. If children used two types of verbalizations more than 33% of the items, then children were allocated to a mixed strategy class. More information about the scoring method can be found in earlier papers (e.g., Elliott & Resing, 2015; Resing et al., 2017).

**Figure 2.** Scoring of verbalized strategy use

**Planning time.** Based on the times of placement of the answering pieces, an adapted version of the formula of Kossowska and Nęcka (1994) was used to calculate the proportion of time the problem solver spent on the initial stages of the task (Resing et al., 2012). The formula calculated
the proportion of time spent on the placement of the first two pieces of the puppet because most children have shown to place the head of the puppet first and then continue to complete the rest of the puppet. The planning time was theorized to represent the time taken to analyze the task before providing an answer. More time spent on the initial analysis was thought to represent a more reflective style of answering, which was thought to be more advanced than a more impulsive style of answering (Resing & Elliott, 2011).

**Total time.** The total time in milliseconds for each session was extracted from the log files. The total time was measured from the start of the testing session, when the console started with the welcome sound, until the end of the last item. The total time for the complete testing cycle consisted of the sum of all sessions per child.

**Chi-square automatic interaction detection tree analysis.** As mentioned by Siegler (1987), averaging process data over multiple items can lead to a distorted picture of strategy use. To investigate the factors that contribute to the prediction of a construct, linear analyses are most often used. Nevertheless, the relationships between process measures and outcome variables often contain complex interactions and nonlinear relationships (Dodonova & Dodonov, 2013; Goldhammer et al., 2014; Tenison et al., 2014), for which linear analyses do not provide adequate tools (McArdle, 2014; Ritschard, 2014). Chi-square automatic interaction detection (CHAID) tree analysis was developed to detect and explore interactions and nonlinear relationships. In this tree analysis, the data are split to achieve the maximal difference between groups on the dependent variable. The splitting of the groups is continued until a predetermined stopping criterion is reached. The CHAID method in particular allows for splitting into multiple groups at once, based on a single predictor variable. The resulting groups are called “nodes” (McArdle, 2014). For each predictor, CHAID determines the number of splits that would provide optimal prediction, and the points at which these splits should be cut off. CHAID was based on the chi-square test, using the p value with Bonferroni correction to determine which splits are made. In addition, the minimum number of cases for each split can be determined manually to avoid overfitting the model (Ritschard, 2014).
3.3 Results

To be able to investigate any differences on the posttest level, any differences between the two groups prior to testing should be ruled out. To this end, a one-way analysis of variance (ANOVA) was conducted with the Raven scores as the dependent variable, and the Condition (Pretest/No Pretest) as the independent variable. No significant difference was found between the groups, $F(1, 65) = 0.36, p = .55, \eta^2 = .01$, indicating that any differences between the groups on subsequent measures could be interpreted as a result of the procedure. In addition, a repeated measures ANOVA was used to ensure the training procedure led to increased accuracy from the pretest to the posttest. The analysis showed a significant effect for session, $F(1, 31) = 12.11, p = .002, \eta^2 = .28$, indicating increased accuracy as a result of the training.

The effect of a pretest on posttest performance

First, this article investigated whether the use of a pretest would lead to differences in performance on the posttest. A one-way ANOVA with Posttest accuracy as the dependent variable and Condition as the independent variable was used (Table 2), which revealed no significant differences between the two groups ($p = .29$). In line with our hypothesis, whether or not a child had received a pretest did not lead to any significant differences on posttest accuracy scores. Furthermore, it was expected that there would be no differences between the group that received the pretest and the group that did not receive the pretest on any of the process measures on the series completion posttest. One-way ANOVAs (shown in Table 2) showed no significant differences between the two groups on the Posttest on GAP ($p = .72$), Verbalized strategy use ($p = .75$), or Planning time ($p = .66$). In line with our expectation, there were no significant differences in the level of advancement of the process used to solve the problems on the posttest between the groups as a result of receiving a pretest.
Table 2. Results of the one-way ANOVAs for Accuracy, GAP, verbalized strategy use, and Planning time.

<table>
<thead>
<tr>
<th></th>
<th>F(1, 65)</th>
<th>p</th>
<th>η²</th>
<th>Mean (SD) Pretest (n=32)</th>
<th>No pretest (n=35)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accuracy</td>
<td>1.16</td>
<td>.29</td>
<td>.02</td>
<td>82.50 (10.00)</td>
<td>79.29 (13.95)</td>
</tr>
<tr>
<td>GAP</td>
<td>.13</td>
<td>.72</td>
<td>.00</td>
<td>.81 (.06)</td>
<td>.81 (.10)</td>
</tr>
<tr>
<td>Verbalized strategy use</td>
<td>.11</td>
<td>.75</td>
<td>.00</td>
<td>2.63 (1.85)</td>
<td>2.77 (1.85)</td>
</tr>
<tr>
<td>Planning time</td>
<td>.19</td>
<td>.66</td>
<td>.00</td>
<td>.43 (.08)</td>
<td>.44 (.11)</td>
</tr>
</tbody>
</table>

The effect of a pretest on the number of hints needed during training

Second, we expected no significant difference between children who had and who had not received a pretest on the amount of hints they needed during the training sessions. To investigate this, first, a one-way ANOVA was used, with the Number of hints needed on the first training session as the dependent variable and Condition as the independent variable. No significant difference between the amount of hints needed was found between the groups, $F(1, 65) = 1.54, p = .22, \eta^2 = .02$. In line with our expectations, the group that had not received a pretest did not require significantly more hints than the group that had received a pretest. Similarly, no differences were found on the number of hints needed between both groups during the second training session, $F(1, 65) = 0.40, p = .53, \eta^2 = .01$. To evaluate any differences in the need for hints over the complete training phase, a repeated measures ANOVA was used, with Session (Training 1/Training 2) as the within-subject factor and Condition as the between-subjects factor. A significant main effect was found for Session, $F(1, 65) = 7.65, p = .007, \eta^2 = .11$, but not for Condition, $F(1, 65) = 1.01, p = .31, \eta^2 = .02$. No significant interaction effect was found for Session 3 Condition, $F(1, 65) = 0.78, p = .38, \eta^2 = .01$. No significant differences were found between the group that had received a pretest and the group that had not on the amount of hints needed during the two training sessions.
Time at the different phases and total time

Regarding the time needed to complete the testing cycle, it was expected that the first training would last significantly longer for the children who had not received a pretest than for the children who had received a pretest. To investigate this, a one-way ANOVA was employed with Time taken for Training 1 as the dependent variable and Condition as the independent variable. The results revealed no significant differences between the groups ($p = .83$). On the second training, posttest, and the full testing cycle, we expected no significant differences in the time they took to administer. One-way ANOVAs were conducted on which no significant differences were found between the group that received the pretest and the group that did not receive the pretest (Table 3) on the time needed for Training 2 ($p = .15$), posttest ($p = .06$), or the total testing cycle ($p = .49$). Table 4 contains the average time in minutes per phase. No differences were found between the two groups on any of the testing phases regarding the time it took to administer them.

| Table 3. | Results of the one-way ANOVAs for time spent on the different phases of testing in seconds. |
| --- | --- | --- | --- | --- | --- |
| | $F$ (1, 65) | $p$ | $\eta^2$ | Mean (SD) |
| | Pretest $(n=32)$ | No pretest $(n=35)$ |
| Training 1 $(n=66)$ | .05 | .83 | .00 | 2250 (1323) | 2319 (1200) |
| Training 2 $(n=66)$ | 2.15 | .15 | .03 | 1788 (1002) | 2266 (1569) |
| Posttest $(n=67)$ | 3.63 | .06 | .05 | 1742 (437) | 2440 (2025) |
| Total testing time $(n=62)$ | .49 | .49 | .01 | 7552 (1868) | 7062 (3287) |

| Table 4. | Time in minutes per phase in the group that had received a pretest and the group that had not. |
| --- | --- | --- | --- | --- |
| | Pretest | Training 1 | Training 2 | Posttest |
| Pretest | 34 | 37 | 30 | 29 |
| No pretest | -- | 39 | 38 | 41 |
| Total | 127 | 118 |
Differences in relations

Next, the prediction of posttest accuracy from process measures was expected to be not significantly different in both groups. A multiple regression analysis was used, with Posttest accuracy as the dependent variable and Number of hints needed during the training sessions, Posttest verbalized strategy use, Posttest GAP, and Posttest planning time as the independent variables. To account for the pretest, a split file for condition was used. The regression for the group who had received a pretest (N = 32) indicated that the model explained 65.0% of the variance, \( R^2 = .70, F(4, 27) = 15.39, p < .001 \). The Number of hints needed during the training sessions significantly predicted Posttest accuracy (\( \beta = 2.76, p < .001 \)). Neither Verbalized strategy use on the posttest (\( \beta = .18, p = .15 \)) nor Posttest GAP (\( \beta = .02, p = .90 \)) and Planning time (\( \beta = .16, p = .16 \)) contributed significantly to the prediction of Posttest accuracy. In the group that had not received a pretest (N = 35), the model explained 53.0% of the variance, \( R^2 = .59, F(4, 30) = 10.57, p < .001 \). Here, too, the Number of hints needed during the training sessions significantly predicted Posttest accuracy (\( \beta = 2.62, p < .001 \)), whereas the other three variables, Verbalized strategy use on the posttest (\( \beta = .21, p = .14 \)), Posttest GAP (\( \beta = .17, p = .17 \)), and Posttest planning time (\( \beta = 2.10, p = .47 \)) did not contribute to the prediction of posttest accuracy. In line with our expectation, the prediction of posttest accuracy from process measures was not significantly different for the group that had received a pretest and the group that had not received a pretest. The number of hints children needed during the training sessions was found to be the only predictor for posttest accuracy for both groups.

It was further expected that the pattern of correlations for the posttest would not significantly differ between both groups. Correlations were calculated between Posttest accuracy, Total number of hints needed during training, Verbalized strategy use, GAP, and Planning time to investigate whether there were different factors between both groups that contributed to task success. The results are presented in Table 5. For the group that had received the pretest (N = 32), the Total number of hints was negatively correlated with Posttest accuracy (r = -.81, p ≤ .001). The only process measure that correlated with Posttest accuracy was the Verbalized strategy (r = .46, p ≤ .01). For the group that had not received the pretest (N = 35), the Total number of hints was also negatively correlated with Posttest accuracy (r = -.71, p ≤ .001). Both Verbalized strategy (r = .46, p ≤ .01) and GAP
(r = .35, p ≤ .05) were correlated with Posttest accuracy. However, Fisher’s r-to-z transformations showed that these differences were not significant for Total number of hints (p = .35), Verbalized strategy (p > .99), GAP (p = .26), or Planning time (p = .80). These findings support our expectation that there no significant differences exist between the patterns of relations with internal measures for both groups.

It was also expected that the pattern of correlations with external measures for cognitive functioning (i.e., school performance) would not significantly differ for both groups. Correlations were calculated between Posttest accuracy, Total number of hints, Verbalized strategy, Posttest GAP, Cito math scores, and Cito reading comprehension scores. The results are depicted in Table 5 and showed moderate to high correlations between Posttest accuracy and Cito math and Cito reading comprehension scores, for both the group that had received the pretest and the group that had not received a pretest. For the group that had received a pretest, none of the process measures correlated significantly with Cito math or reading comprehension scores. For the groups that had not received a pretest, GAP was positively correlated to Cito math (r = .42, p ≤ .05). Verbalized strategy and Planning time were not correlated with Cito math, and none of the process measures showed any significant correlations with Cito reading comprehension scores. In line with our expectations, none of the differences in correlations between the group that had received the pretest and the group that had not were significant.
Table 5. Correlations between process measures in the puppet task and Cito scores

<table>
<thead>
<tr>
<th></th>
<th>Accuracy</th>
<th>Number of hints</th>
<th>Verbalized strategy</th>
<th>GAP</th>
<th>Planning time</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Accuracy</strong></td>
<td>Pretest</td>
<td>-.81 **</td>
<td>.46 **</td>
<td>.08</td>
<td>.01</td>
</tr>
<tr>
<td></td>
<td>No pretest</td>
<td>-.71 **</td>
<td>.46 **</td>
<td>.35 *</td>
<td>.07</td>
</tr>
<tr>
<td><strong>Difference</strong></td>
<td></td>
<td></td>
<td></td>
<td>p = .35</td>
<td>p &gt; .99</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>p = .26</td>
<td>p = .80</td>
</tr>
</tbody>
</table>

| **Cito**               |          |                 |                     |         |               |
| **Math**               | Pretest  | .36 *           | -.52 **             | .08     | .14           |
|                        | No pretest | .51 **        | -.37 *              | .09     | .42 *         |
| **Difference**         |          |                 |                     | p = .49 | p = .46       |
|                        |          |                 |                     | p = .97 | p = .23       |
|                        |          |                 |                     | p = .13 |               |

| **Reading comprehension** | Pretest   | .45 *           | -.55 **             | .14     | .34           |
|                          | No pretest | .47 **        | -.38 *              | .20     | .30           |
| **Difference**           |          |                 |                     | p = .93 | p = .39       |
|                          |          |                 |                     | p = .82 | p = .89       |
|                          |          |                 |                     | p = .14 |               |

Last, the results were analyzed on an item level, because an analysis based on averaged scores over multiple items could lead to a distorted picture of the importance of process measures. It was expected that the process measures would significantly contribute to the prediction of item success, as was found in previous research. Whether or not a child received a pretest was expected to play no role in the prediction of posttest item accuracy. To explore the contributing factors to the prediction of item accuracy, a CHAID tree analysis was used. A separate data file was constructed where each item was treated as a separate case (n = 1,188), and both the pre- and the posttest items were added, along with the process measures per item. Posttest accuracy was used as the dependent variable, and Verbalized strategy, GAP, Planning time, Total time, Item number, Phase (Pretest/Posttest), and Condition (Pretest group/No Pretest group) were added as the independent variables. The minimum number of cases per node was set to N = 30, to avoid overfitting.

The resulting tree model can be seen in Figure 3. The first split was made based on the Item, $F(5, 1182) = 41.27$, $p < .001$, and resulted in six nodes. Node 1 (n = 99), which only included Item 1, was further split based on Condition, $F(1, 97) = 5.58$, $p = .020$. On Item 1, the group that had received a pretest scored significantly higher than the group that had not received a pretest.
Figure 3. CHAID tree for the prediction of Accuracy per item
Node 2 \((n = 693)\) included seven items (Items 2, 4, 5, 6, 7, 8, and 9) and was split further based on the GAP scores, \(F(2, 690) = 12.03, p < .001\) into three groups, where lower GAP scores predicted lower accuracy scores. The group with the highest GAP scores (Node 11; \(n = 473\)) was split further based on Verbalized strategies, \(F(1, 471) = 47.84, p < .001\), where noninductive and partial inductive verbalizations predicted lower accuracy, and full inductive verbalized strategy use predicted higher accuracy. Node 3 \((n = 99)\), which only included Item 3, was split based on Verbalized strategy, \(F(1, 97) = 8.15, p = .011\). On this item, noninductive strategies predicted lower accuracy scores, and partial inductive and full inductive strategies predicted higher accuracy scores. For Item 10 (Node 4) and Item 11 (Node 5), the item identity itself was the only predictor, no further splits could be made based on the available variables. Node 6 \((n = 99)\) was split based on the Phase, \(F(1, 97) = 5.09, p = .026\). On Item 12 during the pretest, children had a higher accuracy than on Item 12 during the posttest.

In this analysis, the items themselves were the primary predictor for accuracy. For most of the items, GAP added to the prediction of task success, and, second, Verbalized strategy use added to the prediction of accuracy. Condition was included as a predictor in the model but only for the first item. Item 12 was the only item on which the prediction was based on whether it was the pretest or the posttest.

### 3.4 Discussion

In this study, we sought to examine the effect of a pretest on the problem-solving process in the posttest and any effects a pretest might have on the validity of the results obtained on the posttest. Although many theories were developed on possible adverse effects of a pretest/training/posttest design (e.g., Kim & Willson, 2010; Klauer, 1993; Sijtsma, 1993), to our knowledge, no studies have actively investigated the effects of using a pretest in the format used in this article.

First, this study investigated whether a pretest would lead to different outcomes on the posttest, both quantitatively on the accuracy of answering the posttest items, and qualitatively on the different process measures.
Both on accuracy and on the process measures no differences were found on the posttest between the group that had received a pretest and the group that had not. Contrary to the theories expressed in the literature about possible adverse effects of retesting (Kim & Willson, 2010; Klauer, 1993), our results did not support the notion that test-retest effects would lead to different outcomes on the posttest. Also on the level of the processes used to solve the tasks, there were no differences between the two groups. The result in both groups seems to be the result of the training. In the light of the effect of training, we could conclude that the training in our design successfully compensated for prior experience with the task, in this case in the shape of the pretest. This would be in line with the long-held belief of some that dynamic testing can serve as a more fair method of testing, as it is influenced to a lesser extent by a child’s prior learning experiences (Elliott et al., 2010; Haywood & Lidz, 2007; Resing & Elliott, 2011; Sternberg & Grigorenko, 2002). The graduated prompts training appeared to equalize children’s differences in experience with the task.

In line with our expectations, children who had not received a pretest did not need more hints during the training sessions than the children who had received a pretest. It would seem that the pretest did not have the learning effect that it was theorized to have by some researchers (e.g., Kim & Willson, 2010; Klauer, 1993). The pretest did not appear to produce any learning that led to increased performance or a different need for instruction during training. Alternatively, this may have been because of interference of the practice in the pretest, with the strategies and skills learned during the training phase. As Opfer and Thompson (2008) stated, for some children, the additional practice of a pretest could lead to a further consolidation of strategies that are not optimal for solving the task at hand. This interference might have led to a less profound learning effect from the pretest.

A concern about dynamic testing that has often been voiced, is the additional time it takes to administer, and the costs associated with this. In this light, cutting time by eliminating the pretest might seem like a viable option. The results did not support the idea that eliminating the pretest saves significant time. This may have been a result of the limited number of participants, as the testing cycle on average took approximately 10 min shorter for the group that had not received a pretest.
Regarding the prediction of posttest accuracy and the relationships between accuracy, process measures, and mathematics and reading comprehension, no significant differences were found between children that had and that had not received a pretest. It seems that a pretest does not influence the construct validity of the posttest, but instead the training influences the construct validity and compensates for the experience of a pretest.

On the prediction per item, a different picture emerged. The primary predicting factor was the item identity, followed by the grouping of answer pieces and verbalized strategy use. Whether or not a child had received a pretest was a factor on the first item only, which is most likely caused by familiarity with the task as a result of doing the pretest (Schorno, 2013). The contributing factors to item success were the same for the pretest and the posttest, except for Item 12. This may have been an effect of the posttest item accidentally being more difficult on the posttest or because of a lack of motivation at the end of the testing cycle. On the contributing factors to item success, process measures seemed to be the most valuable predictors. Although this initially would seem to contradict the increase in accuracy from pre- to posttest, it does not. In the model, the increase in accuracy from pre- to posttest was accounted for by the use of more sophisticated problem-solving strategies. This indicates that the process measures that contribute to item success are the same on both the pre- and the posttest. However, training may lead to more effective activation of these problem-solving processes on the posttest. The processes that lead to the correct solution are activated to a lesser extent during the pretest, especially in children with special needs (Hessels et al., 2011).

**Limitations**

Although the results seem to clearly indicate that no difference is found between designs that use a pretest and designs that do not in terms of construct validity, there are some limitations to these findings that should be taken into account. The low number of participants in this study could have influenced the results, leading to a perceived lack of difference that could potentially emerge with a bigger sample size. For a more complete picture of potential effects of a pretest, Kim and Willson (2010) recommended using a Solomon four-group design, in which one group receives
both a pretest and a training, one group receives a pretest and no training, one group receives a training but no pretest, and the last group receives neither a pretest nor a training.

Furthermore, it should be taken into account that this study was performed within a specific domain and cannot readily be generalized to other domains. Klauer (1993) stated that test-retest effects differ between domains. Although he indicated these effects would be more prevalent in fluid intelligence domains such as inductive reasoning, further research should focus on investigating the effects within different domains of testing. A broader age range would further benefit generalizability of the findings.

It should be noted that our results can not readily be generalized to special education contexts because this group tends to use different processes in learning and on solving tasks than typically developing children or may not activate the processes necessary to solve the task (Hessels et al., 2011; Tiekstra et al., 2009). Future research should investigate the effects of using a pretest with these children on posttest results.

**Implications and future recommendation**

The results of our research were fairly consistent and provided no support for the notion that using a pretest would influence the results or processes on the posttest. Furthermore, in our CHAID analysis, no evidence was found that the basic problem-solving processes differed between the pre- and the posttest phase. The debate on whether or not a pretest is necessary might not be determined by the effect of a pretest on the posttest results but instead be a decision based on the questions to be answered and the situation in which the testing takes place.

In educational settings, the use of a pretest may serve as a static measure of performance and may provide a baseline performance, which serves as a context in which the posttest scores can be interpreted. When no pretest is used, the posttest performance can be compared to the group average. However, it is not possible to see if someone can indeed improve his or her learning as a result of training because there is no baseline to improve on. In addition, for some groups such as children with test anxiety, the pretest may serve as an indicator of the problem experienced in school functioning.
As Vogelaar, Bakker, Elliott, and Resing (2017) point out, children with test anxiety show differential progression paths from pre- to posttest compared to children without test anxiety, characterized by a greater gain in accuracy from pre- to posttest. Testing without a pretest could lead testers to overlook the initial low performance of groups like these, leading to a discrepancy between the results of the dynamic tests and the everyday performance of the child as seen by parents and teachers.