Pre-test versus No Pre-test: An investigation into the problem solving processes in a dynamic testing context

J. Veerbeek, M.G.P. Hessels, S. Vogelaar, & W.C.M. Resing

In Journal of Cognitive Education and Psychology, 2017

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 vs not using a pre-test on both the outcome on the post-test and the processes used in solving inductive reasoning tasks in dynamic testing using a graduated prompts training. Sixty-seven 7-8 year old children were assigned to either a group that received a pre-test or a group that did not receive a pre-test, using a randomized blocking procedure. No significant differences were found between both groups of children on post-test accuracy, process measures, number of hints needed during training, amount of time needed for testing, or the prediction of school related measures. This paper concluded that the decision of whether or not a pre-test is necessary should be based on the research question to be answered, as it does not appear to influence post-test results.

Keywords: dynamic testing; pre-test effect; process assessment; graduated prompts; inductive reasoning.

Pre-test versus No Pre-test: An investigation into the problem solving processes in a dynamic testing context

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 about 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, a number of 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 pre-test, and the threats these pose to the validity and reliability of the test (Kim & Willson, 2010; Klauer, 1993; Sijtsma, 1993). The aim of the current research was to investigate the effect of using a pre-test within the inductive reasoning domain, compared to not using such a pre-test. This effect was studied with regards to both the level of quantitative measurement of post-test results and the qualitative measurements by investigating the processes involved in answering the post-test 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 and 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 utilizes a pre-test/training/post-test format. During the pre- and post-test, 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 that 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 post-test 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 pre-test/training/post-test format. During the pre-test and post-test, 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 pre-test/training/post-test design (Klauer, 1993). Criticism on dynamic testing has mainly been focused on the measurement of change between pre-test and post-test (Sternberg & Grigorenko, 2002). Kim and Willson (2010) however, concluded in their evaluation of the use of pre-testing in educational settings and research that, using a stand-alone pre-test could lead to problems with, among other things, the internal validity of testing (Kim & Willson, 2010).

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). As pre-testing can influence scores on all types of constructs, the exact effects of using a pre-test 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 pre-test may cause higher score on the post-test, but may also lead to the measurement of a qualitatively different construct on the post-test. Training may influence the solution process which, as a consequence, may be different from pre-test to post-test. This implies that a different construct or ability may be measured on the post-test compared to the pre-test (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 post-test are very different from those at pre-test, leading to the conclusion that the measured construct is different on both tests (Hessels et al., 2011; Tiekstra, Hessels, & Minnaert, 2009). The effect of pre-testing 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, as 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 upon 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 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, 2014). According to Newell and Simon (1972), 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, 2014). Restructuring the problem space was found to be closely related to performance and transfer, and the way in which the problem is restructured influences 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, 2016; Resing et al., 2012).

Despite its possible advantages, process-oriented dynamic assessment has not always been optimally used in practice, as 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

The current study aimed to investigate the effects of using a pre-test in a dynamic test within the inductive reasoning domain, both on the product measures on the post-test, and on the process measures on the post-test 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 pre-test on the construct validity of the post-test.

Firstly, this paper investigated whether the use of a pre-test would lead to differences in performance on the post-test. The graduated prompts training utilized 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 pre-test. It was therefore expected that no significant differences would be found between the children that did and did not receive a pre-test. 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, or the percentage of time taken for analyzing the task and planning of the answering process on the post-test of a series completion task.

Secondly, we expected that the children who did not get a pre-test would need the same amount of hints on the training sessions, as the pre-test 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 pre-test 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 the whether or not the children received a pre-test, as retesting only seems to improve superficial familiarity with the type of task (Schorno, 2013).

Third, with regard to the time needed to complete the testing cycle, it was expected that the first training would last significantly longer for the children that had not received a pre-test than for the children that had received a pre-test, 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 post-test between the children that had and that had not received a pre-test. 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 that did not receive a pre-test would save time on not having to make the pre-test, the additional time needed on the other phases was expected to compensate for the difference in number of sessions.

Fourth, the prediction of post-test 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, as the graduated prompts training was expected to equalize the factors involved in post-test success and determine the construct validity of the post-test by providing strategies to successfully solve the post-test 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 pre-test was expected to play no role in the prediction of post-test item accuracy.

Method

Participants

The participants for this study were 67 children, 28 boys and 39 girls (M = 7.9 years; SD = .40). 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 due to illness, and were excluded from the sample.

Design and procedure

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

The Raven's 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 back-up 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 minutes and were performed with approximately one week in between sessions.

Table 1 about here

Materials

Raven's Standard Progressive Matrices. The Raven's 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., 2016). In order to find the correct answer, the child first had to identify the relations between the elements of the series. Secondly, 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 pre-test-training-post-test design. The pre-test and post-test contained 12 items and started with an example item. The pre- and post-test items were isomorph; they were designed to have the same level of difficulty and required the same reasoning processes. Within the pre- and post-test, the items were presented in an increasing level of difficulty.

The training phase consisted of two training sessions, which both contained 6 items. The training sessions utilized 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, modelling 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 RFID tags. The console had a 12x12 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 SD cards, which could be used to import the data on a computer for further analysis.

Scholastic achievement (Cito) 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 percent, '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, as these correspond to the lowest 10 percent scores (Janssen et al., 2015; Jolink et al., 2015; Keuning et al., 2015).

Scoring and analyses

After testing, all data was 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 was 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 3 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 8 body parts. As the pre-test and post-test 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 sessions was calculated. Each training session consisted of 6 items, with a maximum of 4 possible hints per item. Per session this amounted to a maximum score of 24 hints.

Grouping of the answer pieces (GAP). 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 is considered to indicate restructuring of the problem through the use of a division of the problem into smaller sub-problems (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 sub-groups 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. For this item, three groups of puppet pieces were distinguished. The first group of puppet pieces consisted of the three body parts, as 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, as a single piece cannot be grouped. The number of groups per item ranged between 2 and 5.

Figure 1 about here

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, (I) non-inductive, (II) incomplete inductive, and (III) full inductive, as depicted in Figure 2. Based on their verbalizations per item, a score was calculated for the full pre- and post-test. This verbalization class was scored on a five 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., 2016).

Figure 2 about here

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, as 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.

CHAID 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 contains complex interactions and non-linear 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). CHAID tree analysis was developed to detect and explore interactions and non-linear relationships. In this tree analysis, the data is split in order to achieve the maximal difference between groups on the dependent variable. The splitting of the groups is continued until a pre-determined 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. Additionally, the minimum number of cases for each split can be determined manually to avoid overfitting the model (Ritschard, 2014).

Results

To be able to investigate any differences on the post-test level, any differences between the two groups prior to testing should be ruled out. To this end, a one-way ANOVA was

conducted with the Raven scores as the dependent variable, and the Condition (pre-test/no pre-test) as the independent variable. No significant difference was found between the groups ($F(1, 65) = .36$, $p = .55$, $\eta^2 = .01$), indicating that any differences between the groups on subsequent measures could be interpreted as a result of the procedure. Additionally, a Repeated Measures ANOVA was used to ensure the training procedure led to increased accuracy from the pre-test to the post-test. 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 pre-test on post-test performance

First, this paper investigated whether the use of a pre-test would lead to differences in performance on the post-test. A one-way ANOVA with Post-test 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 pre-test did not lead to any significant differences on post-test accuracy scores. Furthermore, it was expected that there would be no differences between the group that received the pre-test and the group that did not receive the pre-test on any of the process measures on the series completion post-test. One-way ANOVAs (shown in Table 2) showed no significant differences between the two groups on the post-test 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 post-test between the groups as a result of receiving a pre-test.

Table 2 about here

The effect of a pre-test on the number of hints needed during training

Second, we expected no significant difference between children that had and that 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 pre-test did not require significantly more hints than the group that had received a pre-test. Similarly, no differences were found on the number of hints needed between both groups during the second training session ($F(1, 65) = .40$, $p = .53$, $\eta^2 = .01$). 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 x Condition ($F(1, 65) = .78$, $p = .38$, $\eta^2 = .01$). No significant differences were found between the group that had received a pre-test 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

With regards to the time needed to complete the testing cycle, it was expected that the first training would last significantly longer for the children that had not received a pre-test than for the children that had received a pre-test. 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, post-test, 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 pre-test and the group that did not receive the pre-test (Table 3) on the time needed for Training 2 (p = .15), Post-test (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 & 4 about here

Differences in relations

Next, the prediction of post-test accuracy from process measures was expected to be not significantly different in both groups. A multiple regression analysis was used, with Post-test Accuracy as the dependent variable and Number of hints needed during the training sessions, Post-test Verbalized strategy use, Post-test GAP, and Post-test Planning time as the independent variables. To account for the pre-test, a split file for condition was used. The regression for the group who had received a pre-test (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 Post-test accuracy (β = -.76, p < .001). Neither Verbalized strategy use on the post-test (β = .18, p = .15), nor Post-test GAP (β = .02, p = .90) and Planning time (β = .16, p = .16) contributed significantly to the prediction of post-test accuracy. In the group that had not received a pre-test (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 Post-test accuracy (β = -.62, p < .001), whereas the other three variables, Verbalized strategy use on the post-test (β = .21, p = .14), Post-test GAP (β = .17, p = .17), and Post-test planning time (β = -.10, p = .47) did not contribute to the prediction of post-test accuracy. In line with our expectation, the

prediction of post-test accuracy from process measures was not significantly different for the group that had received a pre-test and the group that had not received a pre-test. The number of hints children needed during the training sessions was found to be the only predictor for post-test accuracy for both groups.

It was further expected that the pattern of correlations for the Post-test would not significantly differ between both groups. Correlations were calculated between Post-test 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 pre-test (N = 32), the Total number of hints was negatively correlated with Post-test accuracy (r = -.81, $p \leq .001$). The only process measure that correlated with Post-test accuracy was the Verbalized strategy (r = .46, $p \leq .01$). For the group that had not received the pre-test (N = 35), the Total number of hints was also negatively correlated with Post-test accuracy (r = -.71, $p \leq .001$). Both Verbalized strategy (r = .46, $p \leq .01$) and GAP (r = .35, $p \leq .05$) were correlated with Post-test 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 Post-test accuracy, Total number of hints, Verbalized strategy, post-test GAP, Cito math scores, and Cito reading comprehension scores. The results are depicted in Table 5, and showed moderate to high correlations between Post-test accuracy and Cito math and Cito reading comprehension scores, for both the group that had received

the pre-test and the group that had not received a pre-test. For the group that had received a pre-test, none of the process measures correlated significantly with Cito math or Reading comprehension scores. For the groups that had not received a pre-test, GAP was positively correlated to Cito math (r = .42, $p \leq .05$). 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 pre-test and the group that had not were significant.

Table 5 about here

Lastly, 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 pre-test was expected to play no role in the prediction of post-test 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=1188), and both the pre-test and post-test items were added, along with the process measures per item. Post-test accuracy was used as the dependent variable, and Verbalized strategy, GAP, Planning time, Total time, Item number, Phase (pre-test/post-test), and Condition (pre-test group/no pre-test group) were added as the independent variables. The minimum number of cases per node was set to N = 30, to avoid overfitting.

Figure 3 about here

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 6 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 pre-test scored significantly higher than the group that had not received a pre-test.

Node 2 (n=693) included 7 items (item 2, 4, 5, 6, 7, 8, 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 non-inductive 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, non-inductive 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 pre-test, children had a higher accuracy than on item 12 during the post-test.

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 secondly, 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 pre-test or the post-test.

Discussion

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

Firstly, this study investigated whether a pre-test would lead to different outcomes on the post-test, both quantitatively on the accuracy of answering the post-test items, and qualitatively on the different process measures. Both on accuracy and on the process measures no differences were found on the post-test between the group that had received a pre-test 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 post-test. 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 pre-test. 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). It would seem that the graduated prompts training can equalize for children's differences in experience with the task.

In line with our expectations, children that had not received a pre-test did not need more hints during the training sessions than the children who had received a pre-test. It would

seem that the pre-test did not have the learning effect that it was theorized to have by some researchers (e.g. Kim & Willson, 2010; Klauer, 1993). The pre-test 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 due to 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 pre-test 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 pre-test.

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 pre-test might seem like a viable option. The results did not support the idea that eliminating the pre-test saves significant time. This may have been a result of the limited number of participants, as the testing cycle on average took approximately ten minutes shorter for the group that had not received a pretest.

With regards to the prediction of post-test 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 pre-test. It seems that a pre-test does not influence the construct validity of the post-test, but instead the training influences the construct validity and compensates for the experience of a pre-test.

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 pre-test was a factor on the first item only, which is most likely caused by familiarity with the task as a result of doing the pre-test (Schorno, 2013). The contributing factors to item success were the same for the pre-test and the post-test, except for item number 12. This may have been an effect of the post-test item

accidentally being more difficult on the post-test, or due to 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-test to post-test, it does not. In the model, the increase in accuracy from pre-test to post-test 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-test and the post-test. However, training may lead to more effective activation of these problem solving processes on the post-test. The processes that lead to the correct solution are activated to a lesser extent during the pre-test, 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 pre-test 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 pre-test, Kim & Willson (2010) recommended using a Solomon four-group design, in which one group receives both a pre-test and a training, one group receives a pre-test and no training, one group receives a training but no pre-test, and the last group receives neither a pre-test 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, as 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 pre-test with these children on post-test results.

Implications and future recommendation

The results of our research were fairly consistent and provided no support for the notion that using a pre-test would influence the results or processes on the post-test. Further, in our CHAID analysis no evidence was found that the basic problem solving processes differed between the pre-test and the post-test phase. The debate on whether or not a pre-test is necessary might not be determined by the effect of a pre-test on the post-test 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 pre-test may serve as a static measure of performance and may provide a baseline performance, which serves as a context in which the post-test scores can be interpreted. When no pre-test is used the post-test performance can be compared to the group average. However, it is not possible to see if someone can indeed improve his learning as a result of training, as there is no baseline to improve upon. Additionally, for some groups such as children with test anxiety, the pre-test may serve as an indicator of the problem experienced in school functioning. As Vogelaar, Bakker, Elliott, and Resing (2016) point out, children with test anxiety show differential progression paths from pre-test to post-test compared to children without test anxiety, characterized by a greater gain in accuracy from pre-test to post-test. Testing without a pre-test 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.

References

Alibali, M. W., Phillips, K. M. O., & Fischer, A. D. (2009). Learning new problem-solving strategies leads to changes in problem representation. *Cognitive Development*, *24*(2), 89–101. http://doi.org/10.1016/j.cogdev.2008.12.005

Campione, J. C., & Brown, A. L. (1978). Toward a theory of intelligence: Contributions from research with retarded children. *Intelligence*, *2*(3), 279–304. http://doi.org/10.1016/0160-2896(78)90020-X

Campione, J. C., & Brown, A. L. (1987). Linking dynamic assessment with school achievement. In C. S. Lidz (Ed.), *Dynamic assessment: An interactional approach to evaluating learning potential* (pp. 82–109). New York: Guilford Press.

Dodonova, Y. A., & Dodonov, Y. S. (2013). Faster on easy items, more accurate on difficult ones: Cognitive ability and performance on a task of varying difficulty. *Intelligence*, *41*(1), 1–10. http://doi.org/10.1016/j.intell.2012.10.003

Elliott, J. G., Grigorenko, E. L., & Resing, W. C. M. (2010). Dynamic assessment. In P. Peterson, E. Baker, & B. McGaw (Eds.), *International Encyclopedia of Education* (Vol. 3, pp. 220–225). Oxford: Elsevier.

Elliott, J. G., & Resing, W. C. M. (2015). Can Intelligence Testing Inform Educational Intervention for Children with Reading Disability? *Journal of Intelligence*, *3*(4), 137–157. http://doi.org/10.3390/jintelligence3040137

Ericsson, K. A. (2003). The acquisition of expert performance as problem solving. In J. E. Davidson & R. J. Sternberg (Eds.), *The psychology of problem solving* (pp. 31–83). New York: Cambridge University Press.

Feldon, D. F. (2010). Do psychology researchers tell it like it is? A microgenetic analysis of research strategies and self-report accuracy along a continuum of expertise. *Instructional Science*, *38*(4), 395–415. http://doi.org/10.1007/s11251-008-9085-2

Ferrara, R. a, Brown, A. L., & Campione, J. C. (1986). Children's learning and transfer of inductive reasoning rules: studies of proximal development. *Child Development*, *57*(5), 1087–1099. http://doi.org/10.2307/1130433

Goldhammer, F., Naumann, J., Stelter, A., Tóth, K., Rölke, H., & Klieme, E. (2014). The time on task effect in reading and problem solving is moderated by task difficulty and skill: Insights from a computer-based large-scale assessment. *Journal of Educational Psychology*, *106*(3), 608–626. http://doi.org/10.1037/a0034716

Haywood, C. H., & Lidz, C. S. (2007). *Dynamic assessment in practice: Clinical and educational applications*. New York: Cambridge University Press.

Hessels-Schlatter, C., & Hessels, M. G. P. (2009). Clarifying some issues in dynamic assessment: Comments on Karpov and Tzuriel. *Journal of Cognitive Education and Psychology*, *8*(3), 246–251. http://doi.org/10.1891/1945

Hessels, M. G. P., Vanderlinden, K., & Rojas, H. (2011). Training effects in dynamic assessment: A pilot study of eye movement as indicator of problem solving behaviour before and after training. *Educational & Child Psychology*, *28*, 101–113.

Hunt, E. (1980). Intelligence as an information-processing concept. *British Journal of Psychology*, *71*, 449–474.

Janssen, J., Hop, M., & Wouda, J. (2015). *Wetenschappelijke verantwoording Rekenen-Wiskunde 3.0 voor groep 4*. Arnhem.

Jolink, A., Tomesen, M., Hilte, M., Weekers, A., & Engelen, R. (2015). *Wetenschappelijke verantwoording Begrijpend lezen 3.0 voor groep 4*. Arnhem.

Karpov, Y. V, & Tzuriel, D. (2009). Dynamic assessment: Progress, problems, and prospects.

Journal of Cognitive Education and Psychology, *8*(1), 228–237. http://doi.org/10.1891/1945

Keuning, J., Boxtel, H. van, Lansink, N., Visser, J., Weekers, A., & Engelen, R. (2015). *Actualiteit en kwaliteit van normen Een werkwijze voor het normeren van een leerlingvolgsysteem*. Arnhem.

Khandelwal, M., & Mazalek, A. (2007). Teaching table: a tangible mentor for pre-k math education. In *1st International Conference on Tangible and Embedded Interaction (TEI'07)* (pp. 191–194). http://doi.org/10.1145/1226969.1227009

Kim, E. S., & Willson, V. L. (2010). Evaluating pretest effects in pre-post studies. *Educational and Psychological Measurement*, *70*(5), 744–759. http://doi.org/10.1177/0013164410366687

Klauer, K. J. (1993). Learning potential testing: The effect of retesting. In J. H. M. Hamers, K. Sijtsma, & A. J. J. M. Ruijssenaars (Eds.), *Learning potential assessment: Theoretical, methodological and practical issues* (pp. 135–152). Amsterdam/Berwyn, PA, PA: Swets & Zeitlinger Inc.

Klauer, K. J., & Phye, G. D. (2008). Inductive reasoning: A training approach. *Review of Educational Research*, *78*(1), 85–123. http://doi.org/10.3102/0034654307313402

Klauer, K. J., Willmes, K., & Phye, G. D. (2002). Inducing inductive reasoning: Does it transfer to fluid intelligence? *Contemporary Educational Psychology*, *27*(1), 1–25. http://doi.org/10.1006/ceps.2001.1079

Kossowska, M., & Nęcka, E. (1994). Do it your own way: Cognitive strategies, intelligence, and personality. *Personality and Individual Differences*, *16*(1), 33–46. http://doi.org/10.1016/0191-8869(94)90108-2

McArdle, J. J. (2014). Exploratory data mining using decision trees in the behavioral sciences. In J. J. McArdle & G. Ritschard (Eds.), *Contemporary issues in exploratory data mining*

in the behavioral sciences (pp. 3–47). New York: Routledge.

Molnár, G., Greiff, S., & Csapó, B. (2013). Inductive reasoning, domain specific and complex problem solving: Relations and development. *Thinking Skills and Creativity*, *9*, 35–45. http://doi.org/10.1016/j.tsc.2013.03.002

Newell, A., & Simon, H. A. (1972). *Human problem solving*. Englewood Cliffs, NJ: Prentice-Hall.

Opfer, J. E., & Thompson, C. A. (2008). The trouble with transfer: Insights from microgenetic changes in the representation of numerical magnitude. *Child Development*, *79*(3), 788–804. http://doi.org/10.1111/j.1467-8624.2008.01158.x

Perret, P. (2015). Children's Inductive Reasoning: Developmental and Educational Perspectives. *Journal of Cognitive Education and Psychology*, *14*(3), 389–408.

Pretz, J. E., Naples, A. J., & Sternberg, R. J. (2003). Recognizing, defining, and representing problems. In J. E. Davidson & R. J. Sternberg (Eds.), *The psychology of problem solving* (pp. 3–30). New York: Cambridge University Press.

Raven, J., Raven, J. C., & Court, J. H. (1998). *Raven manual: Standard progressive matrices*. Oxford: Oxford Psychologists Press.

Resing, W. C. M. (1993). Measuring inductive reasoning skills: The construction of a learning potential. In J. H. M. Hamers, K. Sijtsma, & A. J. J. M. Ruijssenaars (Eds.), *Learning potential assessment: Theoretical, methodological and practical issues* (pp. 219–241). Amsterdam/Berwyn, PA: Swets & Zeitlinger Inc.

Resing, W. C. M. (2000). Assessing the learning potential for inductive reasoning in young children. In C. S. Lidz & J. G. Elliott (Eds.), *Dynamic assessment: Prevailing models and applications* (pp. 229–262). New York: Elsevier.

Resing, W. C. M., & Elliott, J. G. (2011). Dynamic testing with tangible electronics: measuring children's change in strategy use with a series completion task. *The British*

Journal of Educational Psychology, *81*, 579–605. http://doi.org/10.1348/2044-8279.002006

Resing, W. C. M., Touw, K. W. J., Veerbeek, J., & Elliott, J. G. (2016). Progress in the inductive strategy use of children from different ethnic backgrounds: a study employing dynamic testing. *Educational Psychology*, *3410*(April), 0. http://doi.org/10.1080/01443410.2016.1164300

Resing, W. C. M., Tunteler, E., de Jong, F. M., & Bosma, T. (2009). Dynamic testing in indigenous and ethnic minority children. *Learning and Individual Differences*, *19*(4), 445–450. http://doi.org/10.1016/j.lindif.2009.03.006

Resing, W. C. M., Xenidou-Dervou, I., Steijn, W. M. P., & Elliott, J. G. (2012). A "picture" of children's potential for learning: Looking into strategy changes and working memory by dynamic testing. *Learning and Individual Differences*, *22*(1), 144–150. http://doi.org/10.1016/j.lindif.2011.11.002

Richard, J.-F., & Zamani, M. (2003). A problem-solving model as a tool for analyzing adaptive behavior. In R. J. Sternberg, J. Lautrey, & T. I. Lubart (Eds.), *Models of Intelligence* (pp. 213–226). Washington: American Psychological Association.

Ritschard, G. (2014). CHAID and earlier supervised tree methods. In J. J. McArdle & G. Ritschard (Eds.), *Contemporary issues in exploratory data mining in the behavioral sciences* (pp. 48–74). New York: Routledge.

Robertson, S. I. (2001). *Problem solving*. East Sussex, UK: Psychology Press Ltd.

Schorno, S. (2013). *Analyse de l'effet de l'entraînement du raisonnement analogique: utilisation du capteur du mouvement oculaire dans une procédure d'évaluation dynamique [Analysis of the effect of training of analogical reasoning : using eye movement registration in dynamic assessment]*. University of Geneva. Retrieved from https://archive-ouverte.unige.ch/unige:30693

Serious Toys. (2011). Serious Toys. Retrieved from www.serioustoys.com

Siegler, R. S. (1987). The perils of averaging data over strategies: An example from children's addition. *Journal of Experimental Psychology: General*, *116*(3), 250–264. http://doi.org/10.1037/0096-3445.116.3.250

Siegler, R. S. (1996). *Emerging minds: The process of change in children's thinking*. New York: Oxford University Press.

Siegler, R. S. (2007). Cognitive variability. *Developmental Science*, *10*(1), 104–109. http://doi.org/10.1111/j.1467-7687.2007.00571.x

Siemens, G. (2013). Learning analytics: The emergence of a discipline. *American Behavioral Scientist*, *57*(10), 1380–1400. http://doi.org/10.1177/0002764213498851

Sijtsma, K. (1993). Psychometric issues in learning potential assessment. In J. H. M. Hamers, K. Sijtsma, & A. J. J. M. Ruijssenaars (Eds.), *Learning potential assessment: Theoretical, methodological and practical issues* (pp. 175–194). Amsterdam/Berwyn, PA, PA: Swets & Zeitlinger Inc.

Simon, H. A., & Kotovsky, K. (1963). Human acquisition of concepts for sequential patterns. *Psychological Review*, *70*(6), 534–546. http://doi.org/10.1037/h0043901

Sternberg, R. J. (1985). *Beyond IQ: A triarchic theory of human intelligence*. New York: Cambridge University Press.

Sternberg, R. J., & Grigorenko, E. L. (2002). *Dynamic testing: The nature and measurement of learning potential*. New York: Cambridge University Press.

Stevenson, C. E., Heiser, W. J., & Resing, W. C. M. (2016). Dynamic testing: Assessing cognitive potential of children with culturally diverse backgrounds. *Learning and Individual Differences*, *47*, 27–36. http://doi.org/10.1016/j.lindif.2015.12.025

Stevenson, C. E., Hickendorff, M., Resing, W. C. M., Heiser, W. J., & de Boeck, P. A. L. (2013). Explanatory item response modeling of children's change on a dynamic test of

analogical reasoning. *Intelligence*, *41*(3), 157–168. http://doi.org/10.1016/j.intell.2013.01.003

Tenison, C., Fincham, J. M., & Anderson, J. R. (2014). Detecting math problem solving strategies: An investigation into the use of retrospective self-reports, latency and fMRI data. *Neuropsychologia*, *54*, 41–52. http://doi.org/10.1016/j.neuropsychologia.2013.12.011

Tiekstra, M., Hessels, M. G. P., & Minnaert, A. E. M. G. (2009). Learning capacity in adolescents with mild intellectual disabilities. *Psychological Reports*, *105*, 804–814. http://doi.org/10.2466/pr0.105.3.804-814

Verhaegh, J., Fontijn, W. F. J., Aarts, E. H. L., & Resing, W. C. M. (2013). In-game assessment and training of nonverbal cognitive skills using TagTiles. *Personal and Ubiquitous Computing*, *17*(8), 1637–1646. http://doi.org/10.1007/s00779-012-0527-0

Verhaegh, J., Fontijn, W., & Hoonhout, J. (2007). TagTiles: Optimal challenge in educational electronics. In *TEI'07: First International Conference on Tangible and Embedded Interaction* (pp. 187–190). New York: ACM Press. http://doi.org/http://doi.acm.org/10.1145/1226969.1227008

Verhaegh, J., Hoonhout, J., & Fontijn, W. (2007). Effective use of fun with a tangible interactive console. In *Proceedings of the 4th International Symposium on Pervasive Gaming Applications* (pp. 177–178).

Vogelaar, B., Bakker, M., Elliott, J. G., & Resing, W. C. M. (2016). Dynamic testing and test anxiety amongst gifted and average-ability children. *British Journal of Educational Psychology*, 75–89. http://doi.org/10.1111/bjep.12136

Weisberg, R. W. (2014). Toward an integrated theory of insight in problem solving. *Thinking & Reasoning*, *21*(1), 5–39. http://doi.org/10.1080/13546783.2014.886625

Wiedl, K. H., Schöttke, H., Green, M. F., & Nuechterlein, K. H. (2004). Dynamic testing in

schizophrenia: does training change the construct validity of a test? *Schizophrenia Bulletin*, *30*(4), 703–711.

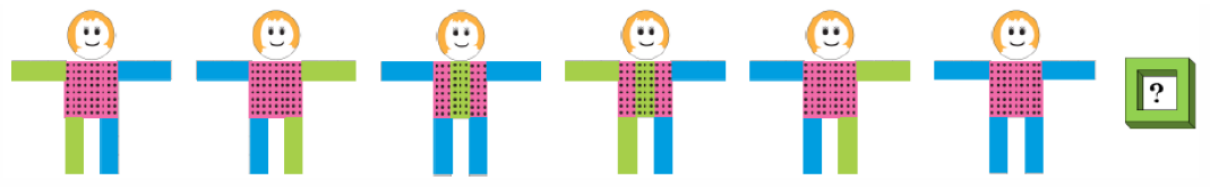

Figure 1. Puppet series completion item

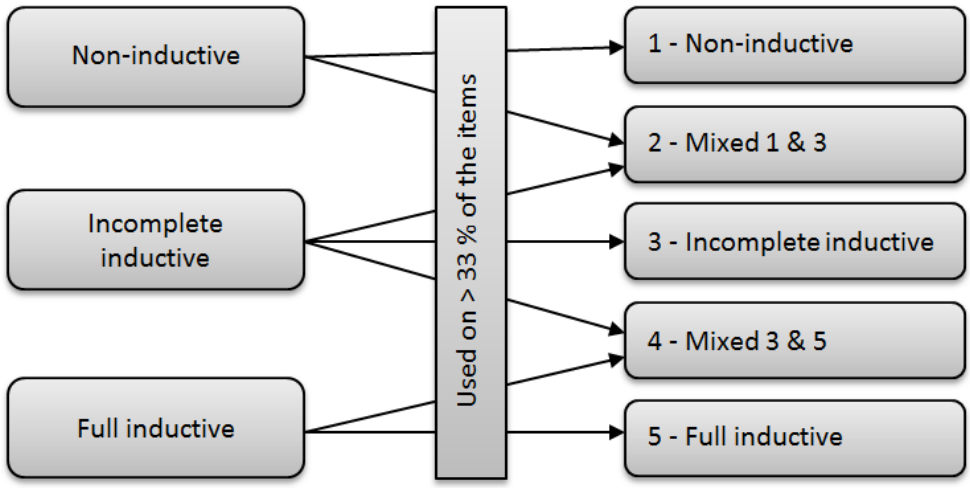

Figure 2. Scoring of verbalized strategy use

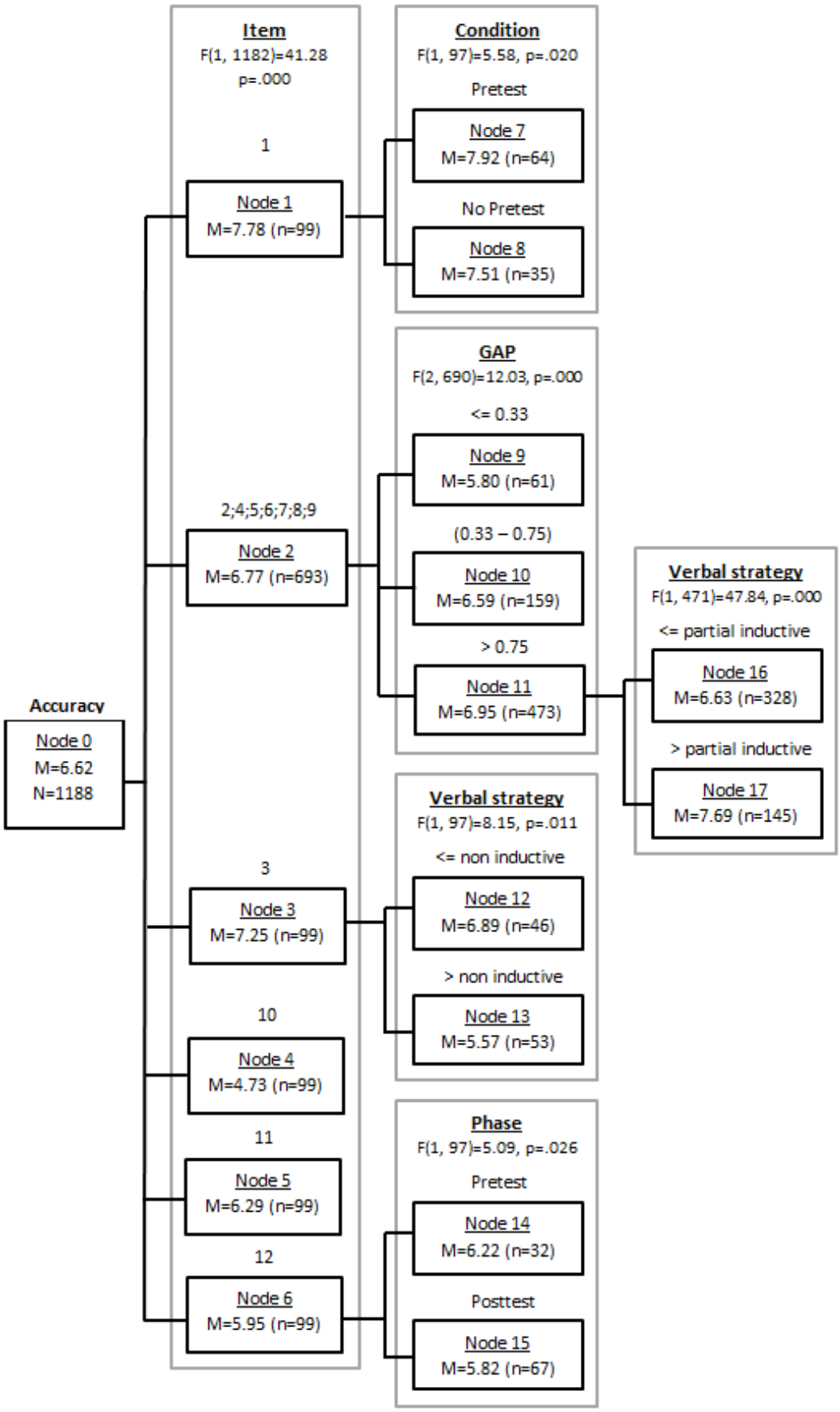

Figure 3. CHAID tree for the prediction of Accuracy per item