

Cover Page



Universiteit Leiden



The handle <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

HIDDEN TREASURES

Uncovering
task solving processes
in dynamic testing



Jochanan Veerbeek

HIDDEN TREASURES

Uncovering
task solving processes
in dynamic testing

Jochanan Veerbeek

© 2019 Jochanan Veerbeek

Typesetting and cover design: Robin Bremer — www.robinbremer.com

Printed by: Ipskamp Printing B.V.

ISBN: 978-94-028-1432-3

Hidden treasures:

Uncovering task solving processes
in dynamic testing

Proefschrift

ter verkrijging van
de graad van Doctor aan de Universiteit Leiden,
op gezag van Rector Magnificus prof.mr. C.J.J.M. Stolker,
volgens besluit van het College voor Promoties
te verdedigen op donderdag 11 april 2019
klokke 16:15 uur

door

Jochanan Veerbeek
geboren te Zwijndrecht
in 1985

Promotor:

Prof.dr. W.C.M. Resing

Co-promotores:

Dr. J. Verhaegh (Userproof)

Dr. B. Vogelaar

Promotiecommissie:

Prof.dr. E.A.M. Crone

Prof.dr. K.H. Wiedl (University of Osnabrück)

Prof.dr. M.G.P. Hessels (University of Geneva)

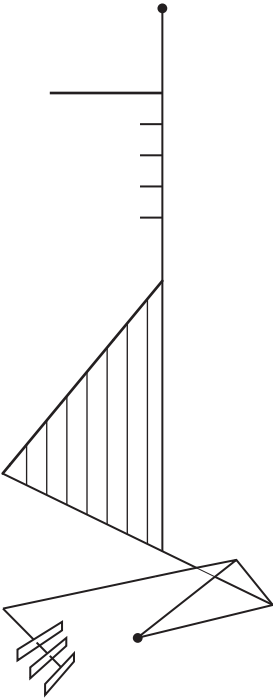
Dr. P.P.M. Hurks (Maastricht University)

Contents

1	General introduction	1
2	Process-oriented measurement using electronic tangibles	11
2.1	Introduction	13
2.2	Method	19
2.3	Results	26
2.4	Discussion	29
3	Pretest versus no pretest: An investigation into the problem-solving processes in a dynamic testing context	35
3.1	Introduction	37
3.2	Method	43
3.3	Results	50
3.4	Discussion	57
4	Process assessment in dynamic testing using electronic tangibles	63
4.1	Introduction	65
4.2	Method	70
4.3	Results	77
4.4	Discussion	88
5	Process-oriented dynamic testing in a complex figure task	97
5.1	Introduction	99
5.2	Method	102
5.3	Results	106
5.4	Discussion	108
6	General discussion	113
	References	125
	Appendix	139
	Summary in Dutch	143
	Curriculum Vitae	151

CHAPTER 1

General introduction



A large part of children's cognitive development takes place in an educational context. Conventional, static, cognitive ability tests are often used as assessment tools when questions arise regarding children's cognitive capabilities or when they do not show the progression in academic achievements their peers do (Elliott, Resing, & Beckmann, 2018). However, these static assessment instruments are said to provide rather limited product-oriented information regarding the potential reasons behind children's success or failure on the test (Elliott, 2000; Elliott, Grigorenko, & Resing, 2010; Haywood & Lidz, 2007; Sternberg & Grigorenko, 2002). They hardly provide information on individual differences regarding the processes children employ to solve (complex) cognitive tasks (Richard & Zamani, 2003). To address these drawbacks, dynamic testing has been developed, which aims to make an estimate of a child's potential for learning on the basis of feedback or intervention methods that lead to successful task solving behavior. Some forms of dynamic testing focus on process assessment, aiming to obtain information regarding the task solving processes children use (Resing, 2013). The main aim of the current thesis was to investigate the information that can be obtained from process-oriented dynamic testing and the specific characteristics of a new method of computer automated process assessment based on children's sequences of solving activities when answering cognitive problems.

Dynamic testing

Dynamic assessment or testing in its broadest definition constitutes an assessment procedure in which a form of feedback is provided to the child (Haywood & Lidz, 2007; Sternberg & Grigorenko, 2002). The various forms of feedback or training that are part of dynamic assessment or dynamic testing methods cover a whole range of instruction procedures (e.g., Haywood & Lidz, 2007; Sternberg & Grigorenko, 2002). Some methods, often referred to as "dynamic assessment" propose mediation that is provided on a highly individual basis (e.g., Feuerstein, Feuerstein, Falik, & Rand, 2002). Other dynamic assessment instruments, often referred to as "dynamic testing", were developed with a focus on standardized test-and-training formats (Sternberg & Grigorenko, 2002).

The graduated prompts method, utilized in this thesis, is a specific form of dynamic testing, which follows a pretest/training/posttest format,

aimed at finding the minimal number of prompts children need to solve tasks (Campione & Brown, 1978; Resing, 1993, 2000). The child is expected to work independently during the pre- and posttest, as help is not provided. During the training phase, graduated prompts approaches follow a standardized format, in which predetermined prompts are provided to children if and when they fail to answer a task correctly (Campione & Brown, 1978; Campione, Brown, Ferrara, Jones, & Steinberg, 1985; Resing, 1993, 2000). If a child is not able to solve a task correctly during training, at first, general, metacognitive prompts are provided. If a child cannot use these prompts to correctly solve the task, more specific, cognitive prompts will be provided. In a third phase, if children are still unable to solve the task correctly having been provided with these prompts already, a scaffolding procedure is offered, in which the process of solving the task is modelled to the child (Resing & Elliott, 2011; Resing, Xenidou-Dervou, Steijn, & Elliott, 2012). Both the number of prompts that children needed during training, and the number of items they could correctly solve at the posttest were found to be reasonably accurate predictors of children's learning potential (Caffrey, Fuchs, & Fuchs, 2008; Resing & Elliott, 2011).

The use of dynamic testing procedures has been shown in past research to provide additional explained variance regarding the prediction of school achievement outcomes when compared to the use of static tests only (Elliott et al., 2018; Grigorenko, 2008). However, computer automated process-oriented dynamic testing with the graduated prompts method, such as the tests used in this thesis, is a relatively new development, in relation to which much is still to be learned.

Process assessment

In psycho-educational research, several different methods for detecting and measuring task solving strategies and processes have been described (Ericsson, 2003; Tenison, Fincham, & Anderson, 2014), although these often represent a trade-off between the accuracy of the measurement, the reactivity it evokes from participants, and the ease of use for the tester (Tenison et al., 2014). In this context, reactivity concerns the influence of changes in the task solving situation on the task solving behavior of the child as a result of the measurement method (Kirk & Ashcraft, 2001; Tenison et al., 2014).

One method of process assessment is that of obtaining verbal reports (explanations) from participants, either during the solving process or immediately afterwards. This is generally seen as a valid method of obtaining data regarding the participants' task solving processes (Ericsson & Simon, 1980; Tenison et al., 2014). However, providing a verbal report during or after task solving might induce some change in the solving processes of participants (Kirk & Ashcraft, 2001; Tenison et al., 2014). Another method is that of time measurement. Recent literature, however, has indicated that the relationship between completion time and accuracy is not necessarily the same for all tasks, but instead might be influenced by the level of complexity of the task (Goldhammer et al., 2014; Scherer, Greiff, & Hautamäki, 2015).

Advances in computer technology, such as easier to use hardware, and more powerful software have created new possibilities for both recording data, and analyzing the actions captured in the data (Khandelwal & Mazalek, 2007; Verhaegh, Fontijn, Aarts, & Resing, 2013; Verhaegh, Hoonhout, & Fontijn, 2007). Despite the clear benefits of computerized testing, it also presents some pitfalls. Especially younger children are thought to experience difficulty using a PC mouse or keyboard interface (Verhaegh et al., 2013), including operations such as “drag-and-drop” procedures on touch-surface tablet interfaces (Price, Jewitt, & Crescenzi, 2015). Tangible User Interfaces (TUIs) provide a more natural modality for children, because these consist of tangible, physical materials, which are outfitted with sensors that enable detecting a child's manipulations of these materials (Verhaegh, Resing, Jacobs, & Fontijn, 2009).

In addition to the user interface, computerized systems need facilities to record data and translate the data into meaningful information or scores through some sort of interpretive step. Usually the data of a child's task solving behavior on a computerized task is saved into log-files (Goldhammer, Naumann, Rölke, Stelter, & Tóth, 2017). As Zoanetti and Griffin (2017) describe, the process of inferring meaningful information from all the actions recorded in the log-files has proven to be challenging. This thesis aimed to provide a rule-based theory-driven method of process assessment which is relatively easy to use. In some of our studies sensor technology was used to detect children's action sequences, which was thought to be an unobtrusive method for recording children's actions (Verhaegh et al., 2013).

Task solving processes

When a child is first presented with a task, he or she attempts to understand the task and how it has to be solved by creating an initial mental representation of the task (Fireman, 1996; Pretz, Naples, & Sternberg, 2003; Robertson, 2001). The efficiency and accuracy of the solving process is thought to be determined, in large part, by the quality of this representation (Robertson, 2001). Therefore, many researchers have argued that the initial representation is a crucial aspect of performance (Hunt, 1980; Pretz et al., 2003). It is often considered to be dependent on the availability of knowledge of the solver and the way this knowledge is organized. The representation changes with new information regarding the task becoming available (Pretz et al., 2003; Robertson, 2001).

A second factor that has been assumed to influence the task solving process is the familiarity of the solver with the type of task. The availability of prior models and experience with solving comparable tasks determines whether the solution can be attained through transfer of a previously used solving strategy, and whether or not the solver has global solving methods available that might lead to the solution (Newell & Simon, 1972; Weisberg, 2014). Newell and Simon (1972) stated that, if the solver has no clear strategy available, general methods such as “means-ends-analysis”, which consists of restructuring the representation through the division of the problem into sub-problems, can be used to find a solution (Robertson, 2001; Weisberg, 2015). Alternatively, strategies can be used to store information surrounding the task more efficiently in the mental representation. Halford and colleagues theorized that these strategies might consist of either combining information into meaningful units, referred to as conceptual chunking, or breaking up a task into simple steps that can be solved sequentially, referred to as the segmentation strategy (Andrews & Halford, 2002; Halford, Wilson, & Phillips, 1998). The task representation serves to organize the information surrounding the task, and is said to determine the strategies that are chosen to try and solve a task. Using inaccurate strategies may often be due to an incorrect representation. Moreover, learning to utilize new strategies has been argued to potentially improve a solver’s representation (Alibali, Phillips, & Fischer, 2009).

In his overlapping waves model of strategy use, Siegler (1996, 2007) described children's strategy use as quite variable, both within and between different trials. Even when the same task is presented repeatedly, children will often use different strategies to solve it. He posed that children choose a strategy that provides a balance between the fastest and most accurate solution. Similarly, accurate performance appears not to be related to any specific strategies, but instead to the ability to adaptively switch between strategies when the task demands it (Hunt, 1980). However, in the longer term, learning is characterized by increased use of more advanced strategies (Siegler & Svetina, 2002). Instability in strategy use is thought to indicate learning (Siegler, 2007).

Grouping of answer pieces (GAP)

A core subject of this thesis was the use of a new process-oriented measure, which was thought to provide information on children's task representations by analyzing how they grouped the different pieces in their sequence of answering. The idea of grouping of answer pieces was based on the basic principle of grouping information together into meaningful "chunks" to reduce the load of the task representation (Halford et al., 1998; Simon, 1974). In this thesis, this was executed using a rule-based algorithm to analyze the sequence of children's actions during the solving process. Pieces of the answer were considered grouped if they were completed in immediate succession of each other. In Appendix I more extensive explanation and operationalization was provided of the principles underlying this measure for the two tasks used in this thesis, an inductive reasoning task involving series completion and a complex figure task.

Inductive reasoning

The current thesis made use of inductive reasoning tasks, which require the solver to infer rules that govern a number of elements, described in terms of their similarities and differences (Klauer & Phye, 2008). A key component to the development of inductive reasoning ability and the level of complexity that children can handle in solving inductive reasoning problems seems to be the number of variables that can be encoded into a single mental

representation (De Koning & Hamers, 1999; Perret, 2015). The complexity of items might be reduced by the use of strategies that increase the number of variables that fit into a single representation (Halford et al., 1998; Halford, Wilson, & Phillips, 2010; Perret, 2015).

In several studies of this thesis, a series completion task was used that was developed and further adapted for computerized administration using TUIs (e.g., Resing & Elliott, 2011; Resing, Touw, Veerbeek, & Elliott, 2017). The task consisted of a figural series completion task, in which a series of puppets is presented, and the child is required to construct the last puppet figure using tangible blocks. To complete the tasks, the child has to identify how the patterns of the puppet pieces change throughout the series and determine the rule that governs these changes (Resing et al., 2012).

For series completion tasks, such as those used in this research, the process of solving series was described by, among others, Sternberg and Gardner (1983) and Sternberg (1985), who divided it into three phases. Initially, the solver has to encode the task into a mental representation, through perceiving the elements of the task and retrieving knowledge that is relevant to interpret them. Then, comparison takes place, which requires inferring the rule, mapping the higher-order rules that are involved, applying this rule and comparing the solution to the available options. Finally, the correct answer has to be justified. Alternatively, Simon and Kotovsky (1963) described the process as starting with scanning the series and attempting to detect the relationships governing the elements through formulating hypotheses. Once a relationship is detected, the length of a full cycle has to be discovered to establish the periodicity of the relationships. Then a rule can be formulated and, once the position of the answer is determined, applied to construct the final answer (Resing & Elliott, 2011).

The sequences of placement while constructing the answer (the correct puppet, consisting of eight separate pieces) were analyzed on grouping of answer pieces, through an automated algorithm that was based on task analysis of all the items in the test. Pieces that went through the same transformations, or shared common characteristics such as color, pattern, or anatomy, were considered grouped if they were placed immediately after each other in the sequence of solving (see Appendix I in the back of this dissertation for a more detailed account).

Complex figure drawing

This thesis also used a dynamic complex figure drawing task (Osterrieth, 1944; Rey, 1941), and reconstructed the grouping of answer pieces measure for use with this type of task. Complex figure drawing tasks can be used to assess visuo-spatial abilities, such as visual perception and construction (Martens, Hurks, & Jolles, 2014), and requires multiple cognitive processes, which include visuo-spatial perception and representation, planning, and working memory. Some researchers underlined the primary role of encoding for developmental changes in complex figure drawing ability, and pointed out that, in general, older children tend to solve the task through more configurational, coherent encoding of the figure (Kirkwood, Weiler, Bernstein, Forbes, & Waber, 2001). The algorithm for detecting grouping of answer pieces was adapted for use with the complex figure task. In this task, it was used to analyze which lines were drawn sequentially to use the configurational elements in structuring the representation of the task. Based on prior analyses of the processes involved in solving complex figure tasks (Kirkwood et al., 2001; Resch, Keulers, Martens, van Heugten, & Hurks, 2018), groups of lines were discerned that were considered useful when sequentially drawn together in time (i.e. drawn as one bigger configuration/cluster of lines, instead of as separate lines).

In this thesis, we explored whether children's grouping behavior would change as a result of training, and how grouping behavior in both the series completion and the complex figure task was related to task performance.

Outline of this thesis

The current thesis aimed to explore and examine aspects of process-oriented measurement in a dynamic testing context. Across all studies, both established process-oriented measures and a new measure were used, which was considered to measure children's grouping behavior in constructed response dynamic tests. The relations between task accuracy and task solving processes were investigated, as well as the effects of training on the sequences in task solving children employed to answer the tasks. The (additional) predictive value of the sequences in task solving for academic performance was topic of study as well.

Chapter 2 seeks to provide more insight into children's task representations by analyzing their sequences of task solving steps. It was examined whether and how children, when solving series completion tasks, grouped the different task manipulatives into meaningful "chunks" (Halford et al., 1998). A tangible user interface (TUI) was used for studying natural task solving behavior by children in combination with computerized monitoring of the task solving steps. The study was executed with 8-year old children from regular school grade 2 classes, and children's GAP scores, verbalizations, completion times, and proportion of time taken on the initial stages of the task were used to evaluate the relationship between the different measures and task accuracy. Additionally, the contributions of the task solving process measures to the prediction of item success were investigated.

Chapter 3 focuses on the effects of using a pretest in dynamic testing with a series completion task. In this study one group of 7-to 8-year old children received a dynamic test including a pretest and a second group of children a dynamic test without a pretest. The effects of utilizing a pretest on accuracy, GAP, verbalizations, and completion time were examined, including the use of a pretest on the number of prompts children need during a graduated prompts training procedure, and on the time children spend on the different stages of testing. Furthermore, this chapter explores the question whether the occurrence of a pretest leads to changes in the relationship between task solving processes in dynamic testing on one hand, and both task performance on the series completion task, and academic performance on math and reading comprehension on the other hand.

Chapter 4 presents 7-to-8-year old children's progression in solving a dynamic test of series completion. The study employed a pretest-posttest control group design with randomized blocking. The effects of a graduated prompts training procedure on the accuracy scores on series completion, as well as shifts in the use of GAP, verbalized strategy use, and completion time were presented. In addition to the effects of training on processes children use while solving series completion tasks, this chapter also sheds a light on the predictive value of these task solving processes in relation to academic performance on math and reading comprehension.

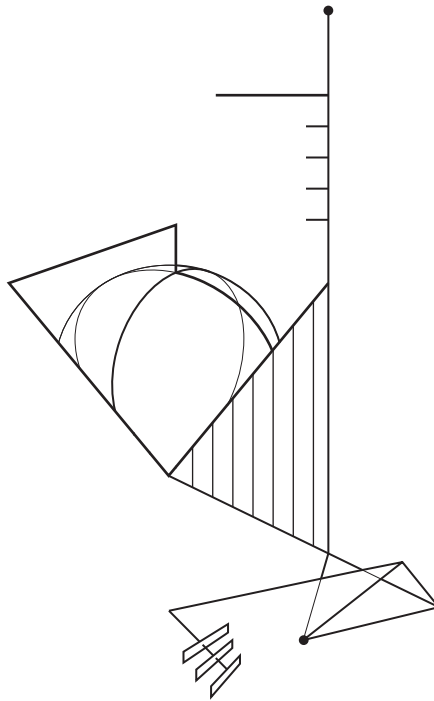
Chapter 5 explores the use of sequential task solving steps (GAP) in a different domain, within a dynamic testing format. A dynamic test of complex figure drawing was administered to children in either a dynamic, graduated

prompts testing condition or a static unguided repeated testing condition. This chapter focuses on the effects and reach of a graduated prompts training on both performance and GAP measures of the dynamic complex figure task.

Chapter 6, the final chapter in this thesis, concludes with an overview of the results of the studies presented in the thesis, in order to synthesize the research findings into a first conclusion regarding the utility of process-oriented dynamic testing. Implications of these findings and recommendations for further research are highlighted.

CHAPTER 2

Process-oriented measurement using electronic tangibles



Jochanan Veerbeek
Janneke Verhaegh
Julian C. Elliott
Wilma C. M. Resing

This chapter was published as: Veerbeek, J., Verhaegh, J., Elliott, J. G., & Resing, W. C. M. (2017). Process-oriented measurement using electronic tangibles. *Journal of Education and Learning*, 6(2), 155-170.

Abstract

This study evaluated a new measure for analyzing the process of children's problem solving in a series completion task. This measure focused on a process that we entitled the Grouping of Answer Pieces (GAP) that was employed to provide information on problem representation and restructuring. The task was conducted using an electronic tangible interface, to allow for both natural manipulation of physical materials by the children, and computer monitoring of the process. The task was administered to 88 primary school children from grade 2 ($M=8.2$ years, $SD=0.50$). GAP was a moderate predictor of accuracy on the series completion task. Averaged over multiple items, GAP, verbalizations and time measures were related to accuracy. On an item level, however, GAP was the only process measure related to item solving success, and this relationship was mediated by item difficulty. Further research is needed to investigate the precise relationship between problem solving and GAP.

2.1 Introduction

Throughout their school careers, children are subjected to a host of assessment procedures that seek to monitor their learning. In school, their cognitive and curricular progress is monitored by achievement tests; outside of the classroom, intelligence tests are sometimes used to assess the child's cognitive ability. However, these instruments have been subject to critique on the grounds that they are unable to provide information on how a child learns, offer few details about why the child failed to learn (Elliott, Grigorenko, & Resing, 2010; Elliott, 2000), and do not yield useful information about what forms of educational intervention might help the child (Elliott, 2000).

An alternative approach to cognitive assessment involves process-oriented measurement (Benson, Hulac, & Bernstein, 2013; Resing & Elliott, 2011). This form of measurement focuses on the *process* of problem solving, instead of, or in addition to, its *products*, and may help to explain why a particular child failed to solve a problem. Studying the operation of cognitive processes within a test situation could potentially yield information that aids the design of subsequent instruction and intervention (Elliott, 2000; Greiff et al., 2013; Van Gog, Kester, Nievelstein, Giesbers, & Paas, 2009).

In line with such reasoning, the general aim of the present study was to examine a new process measure that we called *Grouping of Answer Pieces* (GAP), which was designed to assess problem representation and restructuring. This measure was evaluated on its predictive properties within problem solving both in itself, and in combination with existing measures of the problem-solving process. Finally, this study aimed to evaluate the usefulness of a combination of an electronic tangible interface and dedicated analysis system in process-oriented assessment within a problem-solving framework.

The process of problem solving

How people solve problems has been a major concern within the fields of cognitive psychology and artificial intelligence. While formal intelligence tests have been widely used, these have not proven very successful in helping us understand individual differences in particular problem-solving processes (Richard & Zamani, 2003). The process of problem solving has often been described as cyclical, consisting of (1) problem recognition, (2) definition and representation of the problem, (3) development of a

solution strategy, (4) organization of relevant knowledge, (5) allocation of mental and physical resources, (6) monitoring the progress towards the goal, and finally, (7) evaluation of the accuracy of the solution (Pretz, Naples, & Sternberg, 2003).

The second phase of the problem-solving cycle, the definition and representation phase, has been extensively described by Newell and Simon (1972), who introduced the concept of a problem space, indicating all possible solutions to the problem. According to these authors, problem space can be reduced by breaking down a problem into a set of smaller problems. Here, heuristics can serve as rules that determine how the problem can be divided into a series of smaller problems, leading to a restructuring of the problem space (Pretz et al., 2003). Heuristics are seen as fast rules and procedures for obtaining an answer or decision without the use of an algorithm, and, therefore, do not necessarily always lead to the correct or optimal result (Colman, 2006). Problem-solving strategies and problem representation are thought to influence each other, as both are related to problem-solving performance, as well as to transfer (Alibali, Phillips, & Fischer, 2009).

Measuring the process of problem solving

The literature on the measurement of problem-solving processes has mainly focused on strategy use. A cognitive strategy has been defined by Kossowska and Nęcka (1994) as “*a unique pattern of information-processing which takes place in a problem solving situation*” (p. 33). It is considered to be a vital component of the problem-solving process (Richard & Zamani, 2003; Siegler, 2007) having both an impact on, and being impacted by, learning (Resing, Xenidou-Dervou, Steijn, & Elliott, 2012; Siegler, 2004). Siegler (1996) described high quality learning as not being rigidly connected to one particular strategy, but rather, to one’s ability to adapt strategy use flexibly to the task requirements. In his opinion, variability in strategy use (rather than stable use of one particular strategy) is often indicative of adaptive learning.

Several methods for measuring strategy use are available although each offers a compromise between accuracy of measurement, participant involvement and reactivity, and ease of use (Tenison, Fincham, & Anderson, 2014). Reactivity is understood here as a change in strategy use, as a result of its assessment. In such instances, the observed strategy use is different

to how the participant might otherwise have employed strategies (Kirk & Ashcraft, 2001; Tenison et al., 2014).

Verbal (i.e., oral) reports are widely employed means of assessing strategy use and cognitive processes. Kirk and Ashcraft (2001) suggested that offering a verbal report may influence the natural mental processes of the participant (i.e., inducing reactivity). Such influence could, in theory, lead to improved or reduced performance. Verbal reporting might increase cognitive load demands and, as a result, reduce the mental resources available for the process that is being reported upon. On the other hand, participants might be motivated to work through the task with greater energy and accuracy, as the requirement to offer oral reports might expose their possible errors more publicly. Strategy assessment using verbal reports therefore requires a compromise between report accuracy and participant reactivity (Tenison et al., 2014). Although debate exists about the accuracy and reliability of verbal reports of cognitive processes (e.g., Feldon, 2010), the general consensus is that these provide valid data when obtained under the correct circumstances (Ericsson & Simon, 1980; Tenison et al., 2014). However, as noted above, we must recognize that the requirement to describe their use of cognitive processes may elicit reactivity from the participants, affecting their natural strategy use (Tenison et al., 2014).

Problem-solving speed has generally been presented in the literature as indicative of cognitive ability; the general assumption being that faster is better, although research findings have not unilaterally supported this view (Goldhammer et al., 2014; Scherer, Greiff, & Hautamäki, 2015). High performing participants have tended to be faster than less proficient participants at highly perceptual, automated, low complexity tasks, but they may take more time when tackling more challenging and complex reasoning tasks (Goldhammer et al., 2014). Others (e.g., Kossowska & Nęcka, 1994) have examined the amount of time taken at different stages of task completion. By analyzing the proportion of time spent on the initial stages of the task, an estimate can be made of the portion of time a participant spent on the analysis of the task and their planning of the problem-solving process. Higher performing participants have been found to spend relatively more time than weaker performers on analysis and planning in the initial stages of the task (Kossowska & Nęcka, 1994; Resing & Elliott, 2011; Resing et al., 2012).

Although developments in the field of technology have offered new possibilities for studying strategy use (Ericsson, 2003), advances in process-oriented measurement have yet to lead to widely used practical methods to incorporate process measures into the assessment of learning and cognitive abilities.

Inductive reasoning

Inductive reasoning requires the detection of a rule governing a specific set of elements, and the formulation of a general rule from these elements (Klauer & Phye, 2008), in such a way that reasoning from a particular situation is applied to a general situation (Sternberg, 1985). Inductive reasoning is generally seen as important for learning and transfer (Klauer, Willmes, & Phye, 2002; Resing et al., 2012). Some authors have argued that inductive reasoning is an important component of cross-curricular thinking and learning skills (Greiff et al., 2013; Molnár, Greiff, & Csapó, 2013). A number of different types of tasks are based on the principles of inductive reasoning; these include analogies, series completion, and categorization (Sternberg, 1985). The focus of this paper is on series completion tasks, which require the solver to analyze a series of elements, and complete the series by supplying the missing element(s). Series completion problems exist in a number of shapes and forms, using letters, numbers, geometric figures, colours, etc. Some forms, such as letters and numbers, have a fixed relationship to each other as they have a natural sequence, while others, such as geometric figures and colours do not. Series completion has long been the subject of research, and the processes involved have been described extensively (see Holzman, Pellegrino, & Glaser, 1983; Simon & Kotovsky, 1963; Sternberg, 1985).

Electronic tangibles

In studying the process of problem solving, computers are useful tools for registering test performance. Over the last years, computerized forms of intelligence tests have been introduced. An individual's performance on the Wechsler Intelligence Scale for Children®- Fifth Edition (WISC-V) can be recorded either on paper or using an iPad. Although offering a computerized version, such tools are not designed to measure the process of problem solving; the focus here is still on recording the outcome.

For those who seek to measure problem-solving processes, physical objects offer benefits that the traditional PC or tablet based interfaces would appear to lack. The benefits of physical materials for learning have been advocated by seminal writers such as Piaget, Bruner and Montessori whose theories inspired the development of sets of materials for classroom use. Dienes' multi-base arithmetic blocks (Dienes, 1964), for example, were intended to facilitate comprehension of elementary mathematics by the formation of "qualitative structures", such as the concept of number (e.g., Piaget, 1976). Digitized physical learning materials were introduced by Papert and his student Resnick, who developed so-called "digital manipulatives" (Resnick et al., 1998). Operating across a broader context than schooling alone, the concept of Tangible User Interfaces (TUIs) offers great promise for learning and assessment. TUIs consist of electronically enhanced tangible materials, which permit a seemingly more natural performance by the student and enable the collection and analysis of computer data by an assessor. In contrast to PC or touch-surface tablet applications, where 2 or 3 dimensional representations of objects are typically utilized, electronic tangibles make use of real objects (Verhaegh, Resing, Jacobs, & Fontijn, 2009).

TUIs integrate input and output in physical objects that represent digital information themselves (Ullmer & Ishii, 2000). Graphical User Interfaces (GUIs), such as a PC mouse and a screen, separate input and output modalities, whereas TUIs seamlessly integrate control and representation. Where younger children may experience difficulty performing some actions on touch-surface tablets, such as drag-and-drop procedures (Price, Jewitt, & Crescenzi, 2015), the physical materials that are used in TUIs permit more natural interaction with the interface (Verhaegh et al., 2009), and draw upon the use of a wider range of human skills and abilities such as perception, motor skills and emotion (Dourish, 2004). It has been found that early cognitive development depends mostly on sensory-motor responses (Goswami, 2008), and, thus, the use of tangible interfaces comes naturally to people.

Several possible benefits of TUIs for learning have been described. TUIs are assumed to support playful learning, which enhances children's engagement in scholastic learning tasks. Furthermore, it is likely that they offer a more accessible and direct interface than PC or Mac-based learning applications, and support multisensory learning as well as collaborative play (Manches, O'Malley, & Benford, 2009; Marshall, 2007).

Aims and research questions

The current research concerned an examination of a novel method of process-oriented measurement involving a new measure of strategy use, called *Grouping of Answer Pieces* (GAP). This measure was applied in a series completion construction task with a TUI, an electronic console. The task consisted of puppet figures, which were to be constructed using eight separate pieces. GAP was considered to be indicative of the use of adaptive heuristics, employed to reduce and (re)structure the problem space, and considered to represent the smaller problems that the task had been broken into (Pretz et al., 2003).

GAP in the series completion task was expected to be related to the accuracy of the participants' performance (Richard & Zamani, 2003). However, it was not expected to be related to other measures of strategy use, such as time measures or verbal reports, as these take place during different stages of the problem-solving cycle (Pretz et al., 2003). GAP was thought to add unique predictive and explanatory value to performance on the tangible series completion task, a factor that could hopefully be added to the existing array of measures, such as verbalizations of the problem-solving process, time measurement, and previous inductive reasoning ability.

Variability in strategy use between items was expected to be connected to performance, thus providing additional value in predicting test performance on the tangible series completion task. Participants who displayed greater variability in strategy use between items, were expected to perform better than those showing less variability (Siegler, 1996, 2007).

Finally, we anticipated that, for each item, performance on the tangible series completion task could largely be explained by a combination of measures: initial skill level, GAP, verbal reports, time strategies, and task features. Siegler (1987) has pointed out, however, that averaging data over multiple items can lead to a distorted image of an individual's strategy use, and can result in the loss of valuable information. Additional analysis of each item was expected to prevent the loss of information that could result from using data averaged over multiple items. The study sought to study the relationship between the process measures, task identity, previous ability, and performance on the tangible series completion task. As was found with the relationship between time measures and performance (Goldhammer et al., 2014; Scherer et al., 2015), we expected this relationship to be complex and interactive.

2.2 Method

Participants

The participants in this study were $N=88$ children, 46 boys and 42 girls ($M=8.2$ years; $SD=0.50$), from 4 grade 2 classes of 3 primary schools. The schools were selected on the basis of their willingness to cooperate and were all located in a predominantly middle class area in the Netherlands. Informed consent from the parents was obtained before testing. Three children failed to complete the study due to absence as a result of illness and were excluded from the analyses.

Design and procedure

Participants were first presented with the Raven's Standard Progressive Matrices (Raven, Raven, & Court, 1998). Each participant received his/her own booklet and answer sheet, and was required to complete the matrices independently in the classroom. After the matrices had been completed, each child was taken out of class to work on the tangible series completion task individually on the electronic console. The console provided standardized instruction for all of the participants. An examiner was present at all times to collect and return the children to class, and to oversee the task process. However, they had no role in providing test instructions.

Materials

Raven's Standard Progressive Matrices. The Raven's Standard Progressive Matrices Test (Raven et al., 1998) was used to assess initial cognitive ability. This group test is considered to be a sound indicator of inductive reasoning ability.

Series Completion Task. This study used a schematic-picture inductive reasoning series completion task which was designed specifically for use with the TUI system, although initially designed as a dynamic test incorporating a graduated prompts form of training (e.g., Resing & Elliott, 2011; Resing, Touw, Veerbeek, & Elliott, 2017). The series completion task required the child to detect changes in objects and relationships in a series of puppet figures, and formulate a rule to complete the series. The task, based on the puppet series completion task designed by Resing and Elliott (2011) was intended to provide an indication of each child's inductive reasoning ability. Schematic-picture

tasks such as the tangible series completion task used in this research are seen as more complex than series completion tasks that make use of letters or numbers. While letters and numbers have a fixed relationship to each other, pictures and colours do not. Thus, in order to solve the series, one must first search for repeating combinations of pictures prior to being able to understand the relationships between the elements of the task.

The test consisted of 12 items with increasing levels of difficulty. The test started with an example item. If the child was unable to provide the correct answer on the example item, the console would provide additional explanation to the child, to ensure understanding of what was expected of him/her. Each series item consisted of an initial array of six puppets and the child was asked to complete the sequence by making the seventh puppet. The child had to analyze the changes across successive puppets and find the rule to enable them to complete the task. An example item can be found in Figure 1. Each puppet consisted of 8 separate pieces. The head was a single piece that determined the gender of the puppet (either boy or girl). The 7 pieces that made up the body, arms and legs of the puppet could vary in colour and pattern. There were 4 different colours available (green, blue, pink, and yellow), which could be plain (no pattern), dotted or striped. The design of the eight piece puppet allows for multiple transformations in a series, so the participant is called upon to use a large number of rules in order to complete the tasks.



Figure 1. An example of an item from the puppet series completion task. The child is presented with an array of puppets (drawn on paper) and is asked to construct the puppet that should appear next in the series

Electronic tangible interface. Our study employed an electronic console called “TagTiles” (Serious Toys, 2011), which enabled us to use a computerized environment for assessment, without the issues that manipulation of a touchscreen or mouse bring for young children (Price et al., 2015;

Verhaegh et al., 2009). This incorporated a 12x12 electronic grid, which was equipped with sensors to detect the placement of puppet pieces on its surface, and LEDs which could be programmed to provide visual, brightly coloured feedback. Through its audio output, the console was able to provide appropriate task instructions. The series completion task was completed by placing the pieces on the console. Each contained a unique RFID tag, which enabled the sensors to detect position, timing, and identity of that particular piece on the console's surface. All activity data were automatically saved in log files on SD memory cards. Log files contained rudimentary information about time, identity, and position of pieces placed on the console, and details about the accuracy the answers, per piece and for the item as a whole. The log files that were created by the console were manually cleared of unnecessary data, e.g., accidental movement of pieces, and relevant data were transferred into SPSS for analysis. Wherever possible, missing data, caused by any failure to detect pieces by the console, were retrieved from written records of the child's performance made by the tester during testing.



Figure 2. TagTiles console

Scoring and analysis

Accuracy. Accuracy was used as the primary outcome variable for the series completion task. The scoring was based on the number of correctly placed pieces. As each answer contained eight puppet pieces, the score for each item ranged from 0 to 8. Given that the full test consisted of 12 items, each child's accuracy score could range between 0 and 96. This approach was

expected to enable far greater differentiation than if we had employed more straightforward right or wrong scoring for each item.

Grouping of Answer Pieces (GAP). The concept of grouping in respect of the placement of the answer pieces on the task is similar to the principle of grouping into “chunks” that is widely utilized in memory contexts (Miller, 1994; Simon, 1974). This is based on the structuring of the problem space and the creation of sub-goals or groups by means of adaptive heuristics. As the puppets used in the tangible series completion task are composed of multiple pieces, and the series completion tasks are composed of a number of different relations and transformations, the child’s response sequence is thought to be influenced by the unfolding process of solving the tasks. The subdivisions in the sequence of piece placement were theorized as indicative of the concepts used to define and represent the problem (Pretz et al., 2003).

Task characteristics were used to create sub-groups of puppet pieces that go through the same transformation, or which are grouped on the basis of colour, pattern, or anatomy. Puppet pieces were considered to be grouped if they were placed immediately after each other. First, puppet pieces were numbered, so a sequence could be identified indicating which piece was placed at a particular point in time. The identification numbers of the pieces ranged from one to eight, as follows: (1) head, (2) left arm, (3) right arm, (4) left body, (5) middle body, (6) right body, (7) left leg, and (8) right leg. An example item (Figure 3) illustrates the basic principle of the GAP measure. The Figure includes a sequence of puppets (the task presented to the child) illustrated with some possible sequences of responses. The displayed task consists of three discernible series of transformations. First, the heads go through a series of changes. The second series is that of the arms and legs, which change colour. The third series is that of the body, which stays the same throughout the series. The first sequence displayed in Figure 3 contains no adaptive grouping of answer pieces, as each successive piece placed is one that goes through a different series of transformations. No adaptive groups of puppet pieces were constructed; pieces that transform according to the same rule were not grouped together. Example 2 contains one of the two groups, the “body” group (the 4th, 5th and 6th position, pieces 4, 5 and 6). All the pieces in this group are placed in immediate succession to one other. The “arms+legs” group (pieces 2, 3, 7 and 8) were not placed as a group, as they were interrupted by the body group. Finally, Example 3 contains both

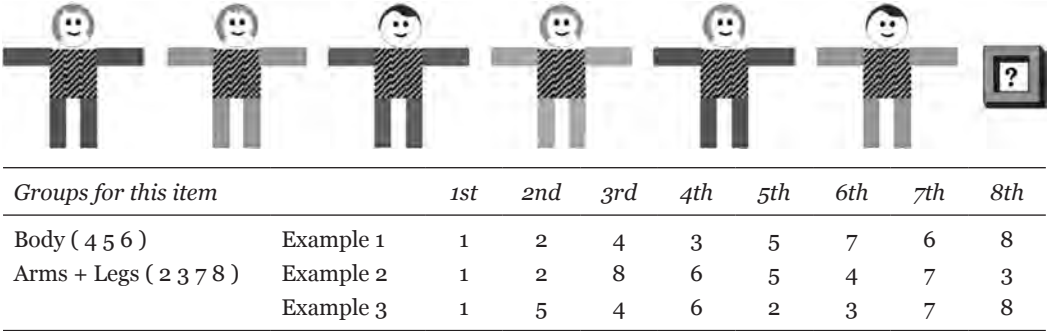


Figure 3. Illustrative responses showing grouping strategy (GAP). Three sequences are provided, with the 1st position being the first piece placed in the sequence, up to the 8th and last piece. The groups for this item are listed in the far left column. Heads are treated as a separate piece and are not included in any of the groups

The adaptive groups of puppet pieces that could be utilised differed between the various items, with the number and type of groups that could be discerned per item ranging between two and five. Some items contained groups that overlapped or contained other smaller groups. For each test item, formulae were written in Microsoft Excel to identify the placement of adaptive groups for a particular item. The GAP score was based on the number of groups laid down by the participant, divided by the number of possible groups for that item.

Verbalized Strategies. After each item was completed, the console’s electronic voice asked “*Why do you think this is the correct puppet?*”. Children’s answers were recorded in writing and by audio recordings, and the verbalizations of their solution strategies were recorded and scored. The scoring system was based on verbalizations that had been found in previous studies, the literature on strategy use in inductive reasoning tasks, and categories used in prior research on reasoning tasks. This resulted in three levels of verbalized strategies, as used by Resing and colleagues (2016), and depicted in Figure 4. In the first group (I), children were able to provide full explanations of all transformations involved in the series, either explicitly or implicitly (e.g., by pointing). The second group (II) contained children who were able to verbalize some transformations in the puppet series, but not all those that were needed to solve the task. The third group (III) consisted of verbalizations that did not provide information relevant to the solution of the task. This scoring was used to analyze each item. The children were then

allocated to 1 of 5 classes of their verbalizations. If children used a single type of verbalization for more than 33% of the items, they were allocated to the corresponding strategy class. If they used two types of verbalizations, both in more than 33% of the items, they were assigned to a mixed strategy class (Figure 4).

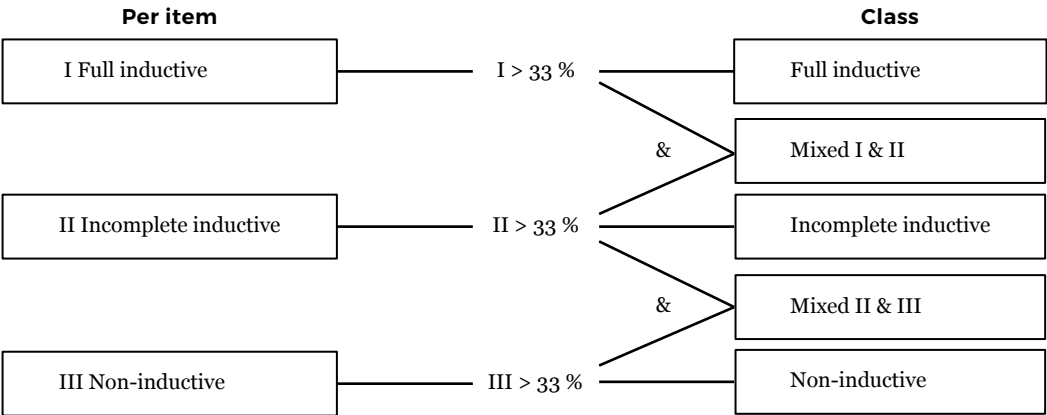


Figure 4. Verbalized strategies. Verbalization is scored for each item. The children were assigned to one of five verbalized strategy classes on the basis of the percentage of items where a particular type of verbalization was provided

Time strategies. Time measures obtained from the log files were the total time for completion of the task (I_{TOTAL} , Figure 5) and the time intervals between the placement of all pieces.

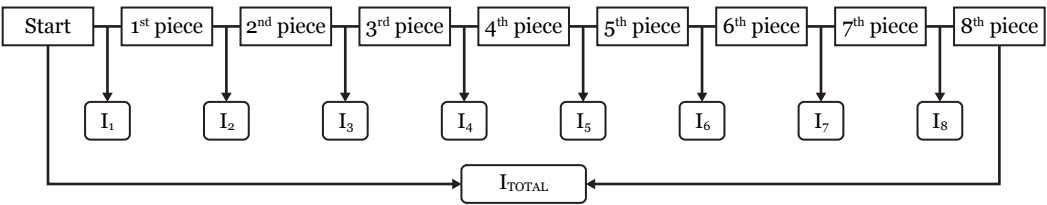


Figure 5. Interval calculation

An adapted version of Kossowska and Nęcka's (1994) formula was used to calculate the proportion of time used on the initial stages of the problem-solving process. These stages are considered to reflect the time taken for analysis and planning of the problem-solving process (Kossowska & Nęcka, 1994; Resing & Elliott, 2011; Resing et al., 2012). The formula calculated the proportion of time used for the placement of the first two pieces, as previous research has shown that many children place the puppet head first and then progress to completing the rest of the puppet. This resulted in the following formula:

$$ThinkingTime = \frac{I_1 + I_2}{I_{TOTAL}}$$

Higher values for thinking time represented more thinking and planning in advance; lower values indicated a more impulsive style of addressing the task (Resing & Elliott, 2011).

Decision Tree Analysis. Classical linear analyses are widely used to investigate contributory predictive factors. However, their usefulness is limited when they are used with data that contain complex interactions and which are non-linear (Ritschard, 2014). Decision Tree Analysis (DTA) contains a number of exploratory techniques aimed at detecting interactions and non-linear relationships within a dataset. The basis of DTA is recursive repartitioning, which involves splitting the data in order to achieve the optimal difference between groups on the outcome variable. This procedure is repeated for each of the splits until an appropriate place to stop is reached. DTAs compare all predictors and search through all possible cut-off points with respect to their effect on the outcome variable. The splitting variable is chosen as a predictor that maximizes the relationship with the outcome variable, at a specific cut-off point (McArdle, 2014). The DTA framework offers multiple techniques and forms of statistical analysis.

We employed the CHAID technique in the present study (McArdle, 2014; Ritschard, 2014). CHAID uses Chi-square analysis as its splitting criterion. The advantage of this approach is that it permits splitting into more than two groups at once for a single predictor. These groups of cases resulting from a split are called "Nodes" in DTA. At each splitting point, CHAID determines the optimal number of splits, and the cut-off points for these splits, for each of the predictors. The p-value of the Chi-square (with a Bonferroni

correction) is used as the criterion to determine the splits. The p-values are sensitive to the number of cases involved, and help avoid any splitting into groups that are too small. The minimum number of cases involved in each split can also be predetermined (Ritschard, 2014), as was done in this study.

2.3 Results

Validating GAP on a test level

Firstly, we examined whether the GAP score was correlated with performance on the task. As expected, we found a positive (albeit moderate) correlation between GAP and Accuracy ($r=.32$, $p=.002$). Additionally, and also as expected, no significant correlations were found for Verbalized strategies ($r=.02$, $p=.88$), TotalTime ($r=-.05$, $p=.65$), or ThinkingTime ($r=.17$, $p=.12$) with GAP. These findings indicate that GAP can be seen as a unique measure of strategy use, unrelated to any previous measures of strategy use.

Multiple regression analysis was used to investigate the prediction of Accuracy in completing the task. Multiple models were tested, and the results are depicted in Table 1. In the first model, Accuracy was used as the dependent variable and GAP was used as the independent variable. As the sole predictor, GAP explained 9.0% of the variance in Accuracy. In Model 2, Verbalization, TotalTime and ThinkingTime were added to the model as independent variables. This model explained 27.0 % of the variance in Accuracy. GAP, Verbalization and ThinkingTime were found to be significant predictors of Accuracy. In contrast, TotalTime was not a significant predictor. The final model (Model 3) contained Raven scores as an independent variable, along with the independent variables used in the previous model. This model explained 30.7% of the variance in Accuracy, with GAP, Verbalization and Raven scores as significant predictors for Accuracy. Neither TotalTime nor ThinkingTime were significant predictors of Accuracy. These findings were in line with our expectations that GAP would add unique predictive value to the available measures.

Table 1. Regression analysis for accuracy on the puppet task

Variable	Model 1			Model 2			Model 3		
	<i>B</i>	<i>SE B</i>	β	<i>B</i>	<i>SE B</i>	β	<i>B</i>	<i>SE B</i>	β
Constant	54.87	7.38		24.32	9.49		19.95	9.43	
GAP	33.72	10.91	.32**	29.83	10.02	.28**	26.40	9.88	.25**
Verbalization				3.11	1.07	.27**	3.21	1.04	.28**
TotalTime				0.00	0.00	.19	0.00	0.00	.16
ThinkingTime				40.84	17.00	.24*	29.20	17.31	.17
Raven scores							0.36	0.15	.23*
R^2		.10			.30			.35	
<i>F</i> for change in R^2		9.55**			7.96**			5.39*	

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

Variability in strategy use

The second hypothesis reflected our expectation that variability in strategy use would be indicative of superior performance on the puppet task, and also, of overall learning. To investigate this, we calculated the variance within each participant's strategy use across all of the items. A multiple regression analysis was used with Accuracy on the task as the dependent variable, and variance in GAP, Verbalization, TotalTime and ThinkingTime as the independent variables. The results are presented in Table 2. This analysis yielded a model that explained 15.1% of the variance in Accuracy. A closer look at the model identified variance in GAP and variance in Verbalization as significant predictors of Accuracy, although it should be noted that GAP was negatively related to Accuracy. Although this was partly in line with our expectations, we had not anticipated the finding that more variability in GAP would lead to less accuracy on the task itself.

Table 2. Regression analysis with variability in strategy use.

Variable	Accuracy		
	<i>B</i>	<i>SE B</i>	β
Constant	76.79	4.86	
GAP	-72.70	34.69	-.21*
Verbalization	31.17	10.35	.31**
TotalTime	-2.06E-10	0.00	-.01
ThinkingTime	163.25	150.18	.12
<i>R</i> ²		.19	
<i>F</i> for change in <i>R</i> ²		4.23**	

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

Validating GAP on an item level

The relationship between the number of correctly placed pieces on the puppet task items, the task characteristics, and the process measures was expected to be complex and interactive. To investigate this, a file was created in which each individual item for a particular participant was handled as a separate case (*N*=1056). For predicting Accuracy on each item (the number of body parts correctly placed out of eight), a Classification Tree was generated, using the CHAID method. The results of this analysis are displayed in Figure 6. Accuracy was used as the dependent variable, and the Item number (the number of the item on the tangible series completion task), scores on the Raven’s progressive matrices, TotalTime, ThinkingTime, GAP and Verbalization were used as the independent variables. The minimum number of cases per node was set to *N*=50.

For Items 4, 6, 9, and 10, the item number was the only factor to explain Accuracy. For Items 1 and 3, Accuracy was further explained by the Raven scores. For Item 2, 5, 7, 8, 11 and 12, GAP was added as an indicator of task success. Higher GAP values seemed to be predictive of higher Accuracy, although Node 10 is an exception to this. Node 10 was split on the basis of Raven’s scores, with higher scores predictive of higher Accuracy. TotalTime, ThinkingTime, and Verbalization failed to offer a significant improvement of the model. At the item level, the best predictors of Accuracy were found to be task characteristics, previous ability (Raven’s scores), and GAP.

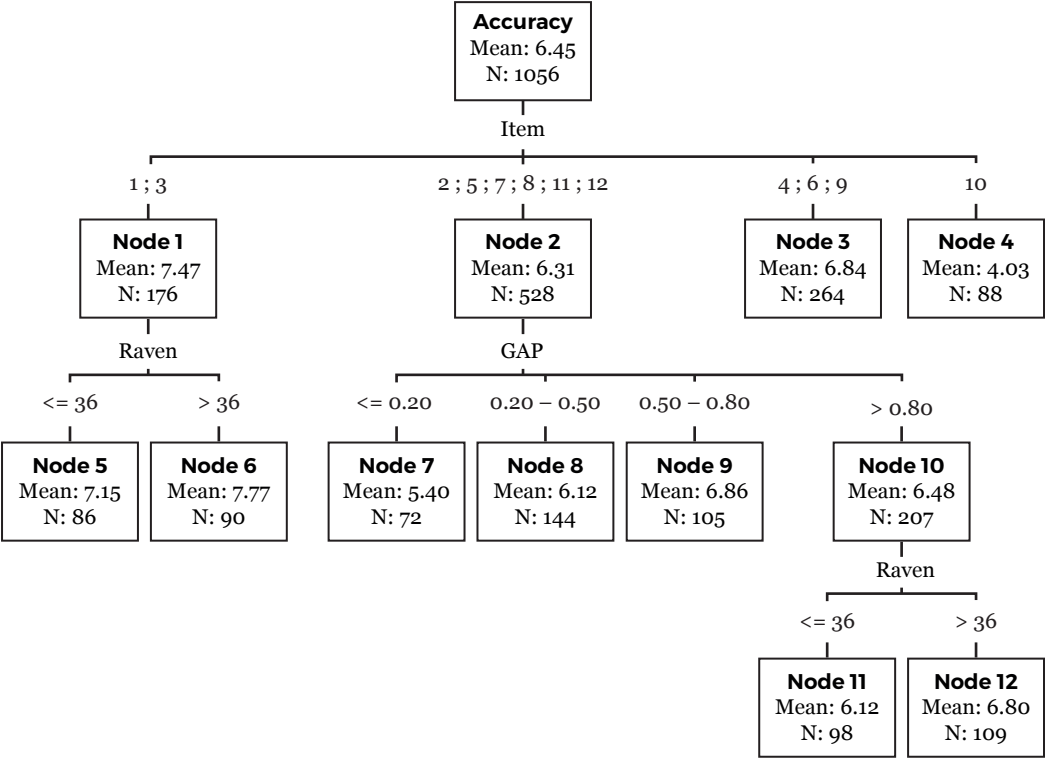


Figure 6. CHAID tree for the prediction of accuracy per item

2.4 Discussion

The aim of this paper was to gain greater understanding of how we can identify and assess the operation of problem-solving processes in young children. To assist in achieving this aim, we utilized a sophisticated assessment tool incorporating electronic tangible technology.

The use of the electronic tangible TagTiles console made it possible to observe and analyze children’s problem-solving strategies in significant detail. GAP appeared to be moderately related to accuracy on the tangible series completion task, as were the previously available measures of strategy use, with the exception of TotalTime. This latter finding may be a consequence of the level of difficulty of the task, as time on task has been shown to be moderated by task difficulty (Goldhammer et al., 2014). The lack of a relation with the

other measures suggests that GAP is a measure that can provide unique information about the process of an individual's problem solving.

Our findings regarding variability in strategy use showed a more complex picture. In line with Siegler's (1996) theory, intra-individual variability in both ThinkingTime and verbalizations was positively related to performance on the task. Variability in GAP was negatively related to accuracy; presumably more stable use of grouping is indicative of greater accuracy in solving the task. If so, GAP appears to function differently to the other strategies. If GAP is more related to general task structuring, as we suspect, it would be a relatively constant style of approaching the process of problem solving, rather than a choice of strategy application. This would make the use of GAP less dependent on particular task content. As GAP is assumed to be related to problem representation (Pretz et al., 2003; Richard & Zamani, 2003), it may be a metacomponent of problem solving. Metacomponents, such as problem recognition, definition, and representation, were described by Sternberg (1985) as executive processes that guide the problem-solving processes, and were expected to be general across cognitive problem-solving activities (Pretz et al., 2003; Sternberg, 1985). As such, GAP may provide more general information on a child's problem-solving processes than previously available process measures, such as verbalizations and time measures, which were found to be more variable.

We were somewhat surprised to discover that the Items themselves were identified as the primary factor for task success as, at least superficially, these seem fairly similar in the characteristics and processes used. However, the results showed that task characteristics were important here. This finding builds upon the work of Goldhammer and colleagues (2014) who found that the relationship between time on task and performance was moderated by task difficulty. In other words, task difficulty is a key factor in determining the need for adequate strategy use. Analysis of the items in the present study indicated this to be true for our task. On the items where there was high average task success (in other words, where the item was found to be relatively easy), the item identity itself provided the best explanation for task performance. As Klauer and Phye (2008) have proposed, experts on a task may use less sophisticated strategies, requiring less time and effort, on easier items. Thus, the use of less sophisticated strategies on easier items may prove more efficient (Siegler, 1996). It was also seen that no single process-

oriented measure provided an explanation for all items. This is in line with Siegler's (1996) notion that averaging data over multiple items might lead to the oversight of important information and may lead to an oversimplification of models of strategy use. GAP provided the single best explanation of all process-oriented measures for performance, as it was the only process-oriented measure included in the model. However, it was not a predictor for accuracy in all of the tangible series completion items. The relationship between strategy use and performance appeared to be moderated by task characteristics (Dodonova & Dodonov, 2013; Goldhammer et al., 2014; Tenison et al., 2014), and might be distorted by the use of linear analyses.

Although our two time measures offered additional explanatory value, no analysis or model included them both. Neither did they complement each other in the prediction of task performance, although ThinkingTime proved to have greater explanatory value. The finding that thinking and planning time provided more information than did total time is in line with the results of a study by Resing and Elliott (2011), who found that total completion time failed to discriminate between their trained and untrained participants.

Process-Oriented Measurement

Although process-oriented measurement yielded some additional explanatory value, a number of complexities emerged. In line with previous research, we found that process-oriented measures are dependent on task characteristics and do not show a unilateral relationship with performance (Goldhammer et al., 2014; Scherer et al., 2015). While process measures may provide additional information for the more difficult items, their relationship with easier items is unclear.

Process-oriented measurement is often labelled as strategy use in the literature. Such inconsistency in the use of terms may lead to confusion. Time measures, verbalizations and actions (GAP or otherwise) have all been labelled as strategy use in previous studies, but they all take place in, and over, different stages of the problem-solving cycle (Figure 7). Although the problem-solving cycle is not necessarily followed in a straightforward, linear fashion (Pretz et al., 2003), the different stages in the problem-solving cycle, where strategy measurement can take place, may explain why the different measures of strategy use were found to be unrelated.

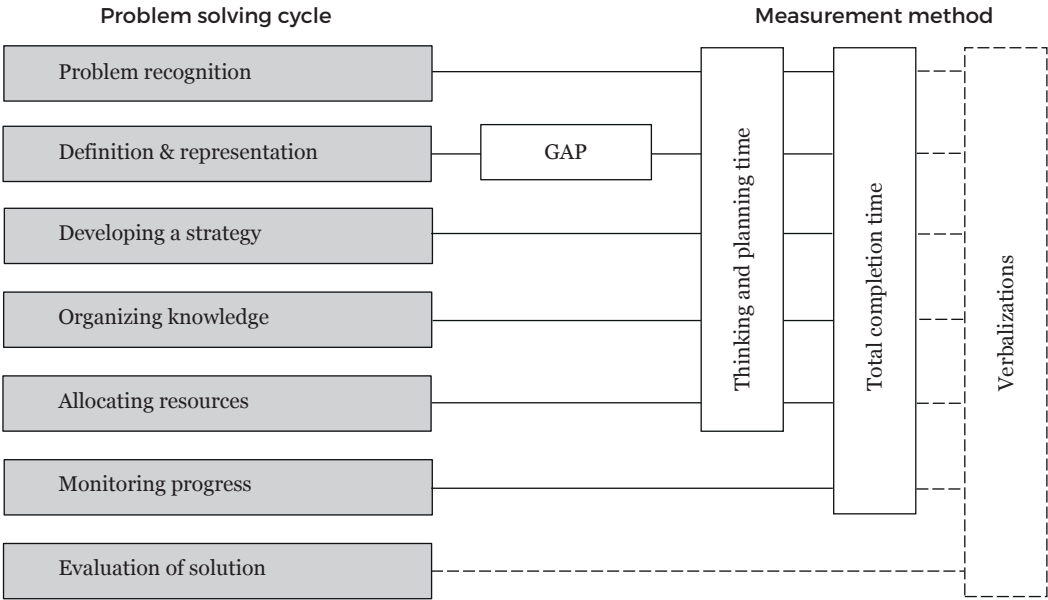


Figure 7. Different process-oriented measures with respect to the phases of the problem solving cycle (Pretz et al., 2003)

GAP has a particular place in the problem-solving cycle, as it may be influenced by the phases between representation and taking physical action. Thinking time represents the time taken on the initial phases of the task, up to the point that physical action begins. Total completion time represents the time taken throughout the whole process, with the exception of the evaluation phase. Verbalizations about strategy use may concern any of the problem-solving phases, but would rarely cover all of these. The particular focus of such accounts will most likely be influenced by the type of verbalization measurement (Tenison et al., 2014), the accompanying instructions (Ericsson & Simon, 1980) and the participant’s willingness and ability to reflect, report and elaborate upon his or her cognitive processes. Taking this into account, a more precise differentiation in the use of terminology could be desirable. We would suggest reserving the term *strategy* for any domain specific procedure, such as the math-specific strategies described by Siegler (1996). Differences in procedures aimed at general structuring of the problem-solving process,

such as the time taken for planning and analysis (Kossowska & Nęcka, 1994), or the grouping performed during problem solving, may be more accurately termed as *process structures*.

Although these findings offer some promising results with regard to process-oriented measurement using electronic tangibles, and the GAP measure in particular, caution is advised in interpreting these findings. As this research solely employed a tangible series completion task, our results cannot be generalized to other domains of problem solving. Even within the field of series completion, the particular task used in the present study cannot be readily generalized to other series completion tasks as these may not contain multiple transformations, or may be more domain-bound by the use of letters or numbers (Resing & Elliott, 2011).

It is also important to note that our sample size was rather small and spanned a narrow age range, thus further limiting generalizability. Clearly, more research is needed, utilizing larger and more diverse samples, in order to enable general statements to be made about the value of process-oriented measurement in education and clinical settings.

Although we found a relationship between performance and process measures in our particular test domain, its nature remains unclear. Future research should determine whether process training can be successfully employed to improve children's intellectual performance.

Conclusions and recommendations for future research

In summary, our GAP procedure appears to offer additional and unique explanatory value to the field of process-oriented measurement. GAP can be measured and interpreted by technological systems such as that employed in the present study, and thus used in computerized testing environments where no examiner is present. Such a facility also makes the measure particularly suitable for analyses of classroom based problem solving. Although theoretically, it would be possible to measure and analyze such processes without the use of technology, this would not be recommended as the adaptive groups differ between items. Manual analysis would be prone to mistakes and oversights, and would be very time consuming. Tangible user interfaces, in contrast, make it possible for educators to provide individualized forms of adaptive instruction, based on the real-time activities/responses of the child.

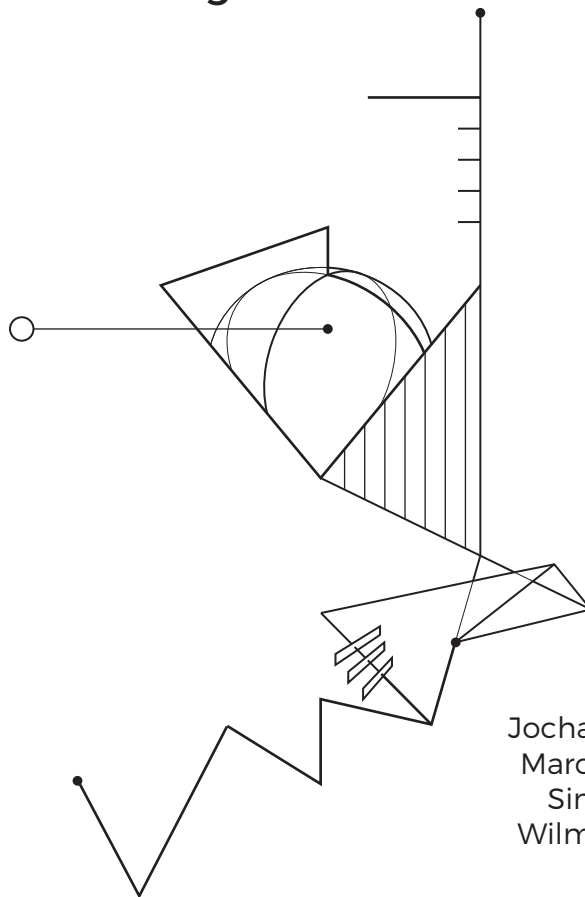
Further research is needed to determine how to interpret the values derived from the obtained measures. In this respect, classification and regression trees may provide a more informative view than traditional analyses, as they are able to handle more complex and interactive data. This will enable researchers to take into account the interaction between item characteristics and the various process measures.

As GAP is based on the subdivision of tasks into sub goals, it is not possible to use this measure with all tasks. Future research should be aimed at identifying a range of diverse tasks that can enable measurement of this kind. The GAP measure does provide an opportunity for analyzing the problem-solving process of participants who for some reason are not able to provide (reliable) verbalizations of their strategies, such as those with specific language difficulties, certain children from ethnic minorities, etc. As the measurement is unobtrusive, there is no risk of participant reactivity. Additionally, tangible user interfaces offer the potential of providing individuals or groups of children with adaptive scaffolds, based on their differing responses to challenging classroom material. In that way, individualized training and assessment of the process of problem solving in education may come within reach.

CHAPTER 3

Pretest versus no pretest

An investigation into the
problem-solving processes in a
dynamic testing context



Jochanan Veerbeek
Marco G. P. Hessels
Simone Vogelaar
Wilma C. M. Resing

This chapter was published as: Veerbeek, J., Hessels, M. G. P., Vogelaar, S., & Resing, W. C. M. (2017). Pretest versus no pretest: An investigation into the problem-solving processes in a dynamic testing context. *Journal of Cognitive Education and Psychology*, 16(3), 260-280.

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.

1
2
3
4
5
6

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 (Campion & 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 welldefined 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 1. Overview of procedures for Treatment and Control group.

	Raven	Pretest	Training 1	Training 2	Posttest
Pretest	X	X	X	X	X
No Pretest	X	No pretest	X	X	X

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

1
2
3
4
5
6

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, Hilde, 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.

1
2
3
4
5
6

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.

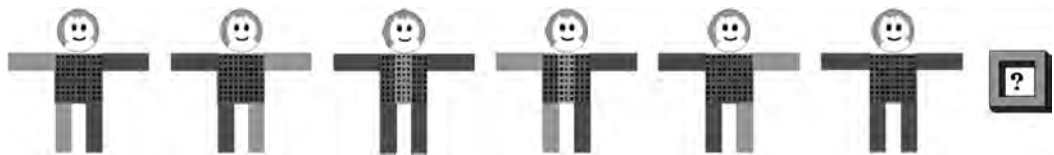


Figure 1. Puppet series completion item

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).

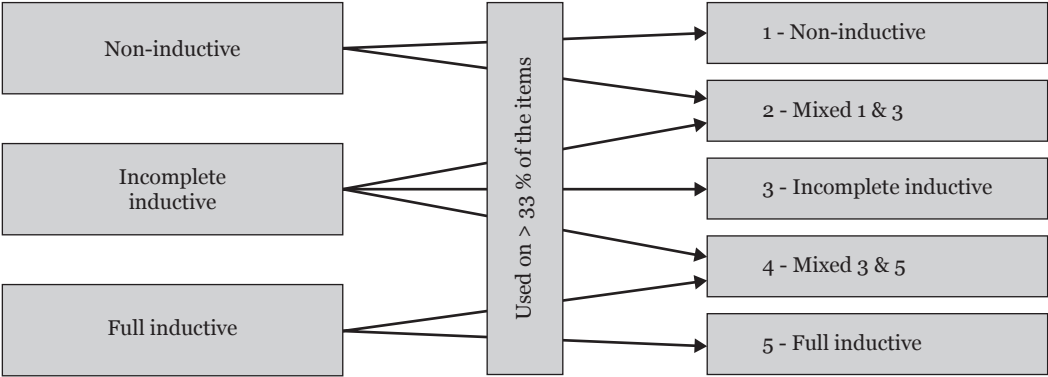


Figure 2. Scoring of verbalized strategy use

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

1
2
3
4
5
6

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.

1
2
3
4
5
6

Table 2. Results of the one-way ANOVAs for Accuracy, GAP, verbalized strategy use, and Planning time.

	<i>F</i> (1, 65)	<i>p</i>	η^2	<i>Mean (SD)</i>			
				Pretest (<i>n</i> =32)		No pretest (<i>n</i> =35)	
Accuracy	1.16	.29	.02	82.50	(10.00)	79.29	(13.95)
GAP	.13	.72	.00	.81	(.06)	.81	(.10)
Verbalized strategy use	.11	.75	.00	2.63	(1.85)	2.77	(1.85)
Planning time	.19	.66	.00	.43	(.08)	.44	(.11)

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.

	<i>F</i> (1, 65)	<i>p</i>	η^2	<i>Mean (SD)</i>			
				Pretest (<i>n</i> =32)		No pretest (<i>n</i> =35)	
Training 1 (<i>n</i> =66)	.05	.83	.00	2250	(1323)	2319	(1200)
Training 2 (<i>n</i> =66)	2.15	.15	.03	1788	(1002)	2266	(1569)
Posttest (<i>n</i> =67)	3.63	.06	.05	1742	(437)	2440	(2025)
Total testing time (<i>n</i> =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	Total
Pretest	34	37	30	29	127
No pretest	--	39	38	41	118

1
2
3
4
5
6

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 ($r = -.81$, $p \leq .001$). The only process measure that correlated with Posttest accuracy was the Verbalized strategy ($r = .46$, $p \leq .01$). For the group that had not received the pretest ($N = 35$), the Total number of hints was also negatively correlated with Posttest accuracy ($r = -.71$, $p \leq .001$). Both Verbalized strategy ($r = .46$, $p \leq .01$) and GAP

($r = .35, p \leq .05$) 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 ($r = .42, p \leq .05$). 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 pretest and the group that had not were significant.

1
2
3
4
5
6

Table 5. Correlations between process measures in the puppet task and Cito scores

		Accuracy	Number of hints	Verbalized strategy	GAP	Planning time
Accuracy	Pretest		-.81 **	.46 **	.08	.01
	No pretest		-.71 **	.46 **	.35 *	.07
	Difference		$p = .35$	$p > .99$	$p = .26$	$p = .80$
Cito						
Math	Pretest	.36 *	-.52 **	.08	.14	-.27
	No pretest	.51 **	-.37 *	.09	.42 *	.12
	Difference	$p = .49$	$p = .46$	$p = .97$	$p = .23$	$p = .13$
Reading comprehension	Pretest	.45 *	-.55 **	.14	.34	-.27
	No pretest	.47 **	-.38 *	.20	.30	.11
	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.

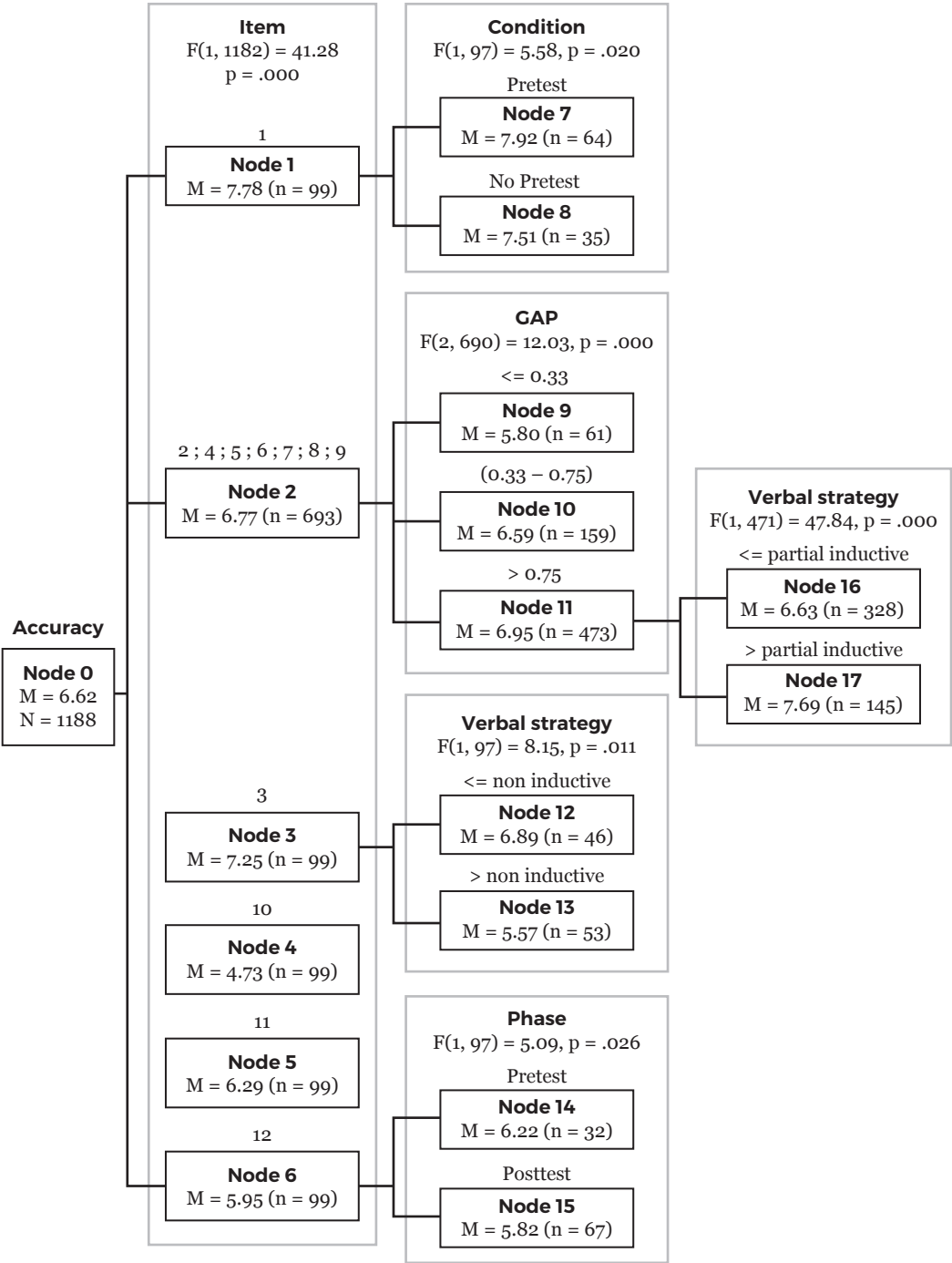


Figure 3. CHAID tree for the prediction of Accuracy per item

1
2
3
4
5
6

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.

1
2
3
4
5
6

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

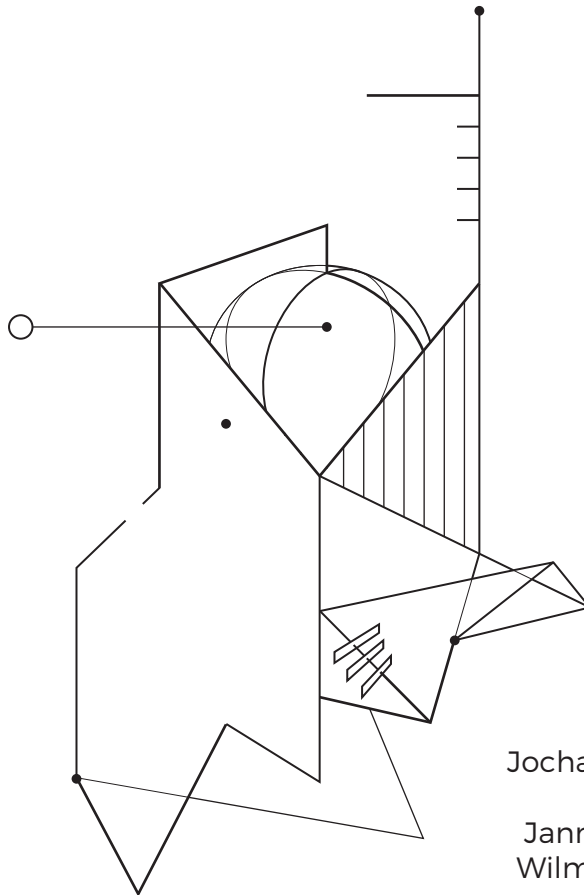
The results of our research were fairly consistent and provided no support for the notion that using a pretest would influence the results or processes on the posttest. Furthermore, in our CHAID analysis, no evidence was found that the basic problem-solving processes differed between the pre- and the 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.

CHAPTER 4

Process assessment in dynamic testing using electronic tangibles



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

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

Abstract

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

4.1 Introduction

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

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

Dynamic testing and graduated prompts procedure

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

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

Inductive reasoning and series completion

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

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

Process-oriented testing

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

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

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

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

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

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

Aims and research questions

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

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

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

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

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

4.2 Method

Participants

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

Design

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

1
2
3
4
5
6

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

Table 1. Overview of procedures for Training and Control group

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

Materials

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

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

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

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

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

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

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

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

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



Figure 1. Example item of the puppet series completion task

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

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

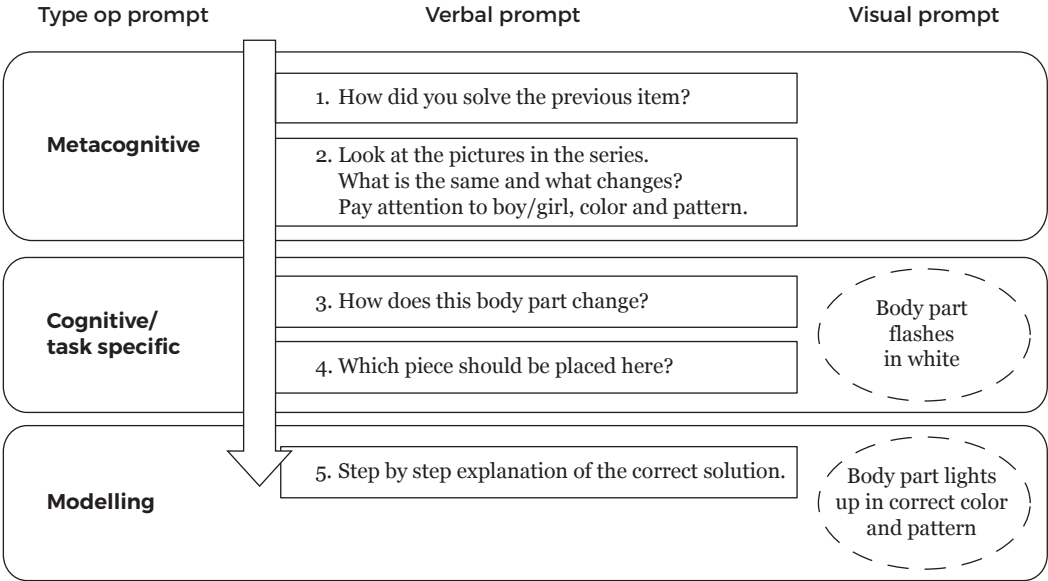


Figure 2. Prompts provided by the electric console during the training procedure of the dynamic test

Scoring

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

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

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

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

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

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

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

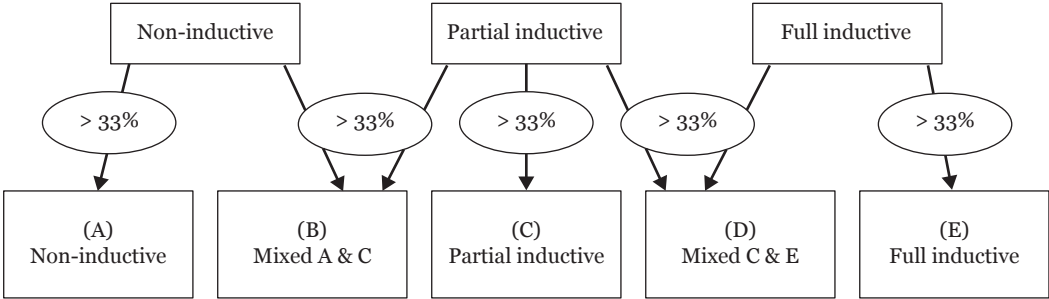


Figure 3. Scoring of verbalized strategy class

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

4.3 Results

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

The effect of training

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

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

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

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

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

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

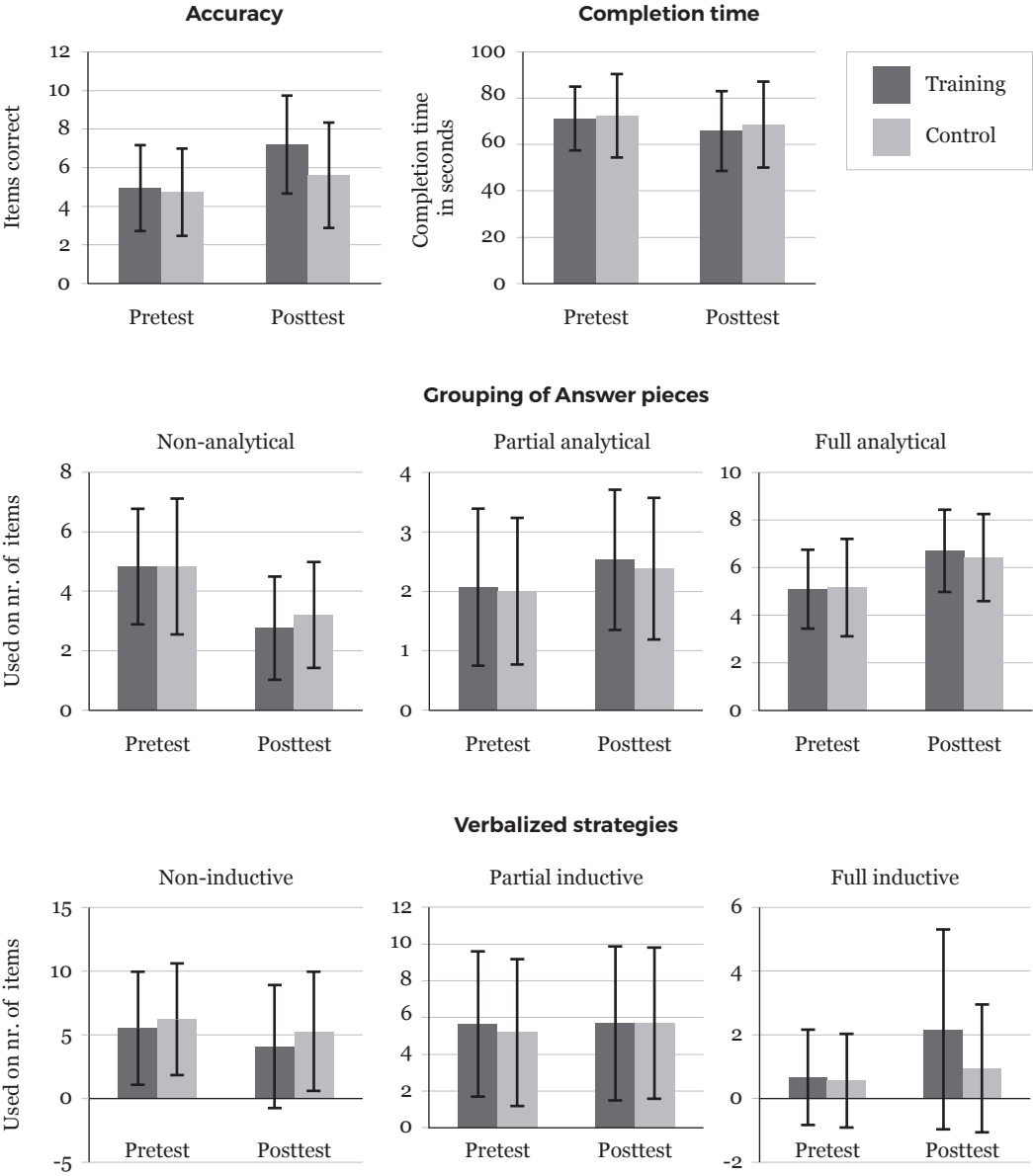


Figure 4. Mean pre- and posttest scores and standards deviations for accuracy, completion time GAP, and verbalized strategies

1
2
3
4
5
6

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

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

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

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

Changes in task solving process over time

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

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

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

		1. Non- analytical	2. Mixed 1 and 3	3. Partial analytical	4. Mixed 3 and 5	5. Full analytical	Missing	Total
<i>Grouping of pieces – Training</i>								
Pretest	Frequency	32	2	40	1	51		126
	Percentage	25.4	1.6	31.7	0.8	40.5		100
Posttest	Frequency	6	0	16	2	102		126
	Percentage	4.8	0.0	12.7	1.6	81.0		100
<i>Grouping of pieces - Control</i>								
Pretest	Frequency	46	1	25	1	54		127
	Percentage	36.2	0.8	19.7	0.8	42.5		100
Posttest	Frequency	18	0	9	3	97		127
	Percentage	14.2	0.0	7.1	2.4	76.4		100
		1. Non- inductive	2. Mixed 1 and 3	3. Partial inductive	4. Mixed 3 and 5	5. Full Inductive	Missing	Total
<i>Verbal explanation – Training</i>								
Pretest	Frequency	54	10	56	4	1	1	126
	Percentage	43.2	8.0	44.8	3.2	0.8		100
Posttest	Frequency	40	7	51	10	18		126
	Percentage	31.7	5.6	40.5	7.9	14.3		100
<i>Verbal explanation - Control</i>								
Pretest	Frequency	57	18	50	1	1		127
	Percentage	44.9	14.2	39.4	0.8	0.8		100
Posttest	Frequency	49	15	51	9	3		127
	Percentage	38.6	11.8	40.2	7.1	2.4		100

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

4.4 Discussion

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

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

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

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

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

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

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

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

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

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

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

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

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

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

APPENDIX A.

Grouping of Answer Pieces, groups per item.

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

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

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

APPENDIX B.
Categories of grouping behavior and verbal strategies.

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

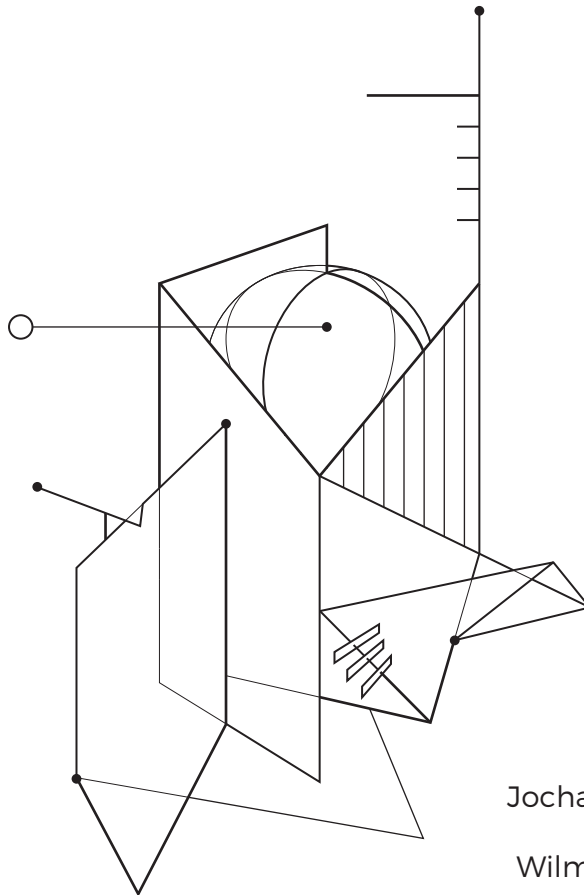
Grouping behavior	Description of category per item
Full analytical	Based on a GAP score of >99% for an item, which indicates adaptive grouping of the puppet parts, based on the transformations in the item (pieces that go through similar transformations are grouped together) and similarity in other characteristics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

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

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

CHAPTER 5

Process-oriented dynamic testing in a complex figure task



Jochanan Veerbeek
Bart Vogelaar
Wilma C. M. Resing

This chapter was submitted for publication: Veerbeek, J., Vogelaar, B., & Resing, W. C. M. (under revision). Process-oriented dynamic testing in a complex figure task. *[Applied Neuropsychology: Child]*

Abstract

Process-oriented dynamic testing aims to investigate the processes children use to solve cognitive tasks, and changes in these processes as a result of training. For the current study, a dynamic complex figure task was constructed, using the graduated prompts approach, to investigate the processes involved in complex figure task solving and changes in these processes after training. A new process oriented measure was developed, which used automated scoring to evaluate children's organization in drawing the figure. Participants were 106 regular primary school children ($M=7.8$ years, $SD=0.42$ years). The graduated prompts training led to significantly more progression in complex figure drawing performance from pretest to posttest, compared to unguided control. The training did not yet lead to transfer from the trained domain to an inductive reasoning task. The training also led to more advanced organization of the figure, and the most advanced category of organization of the figure was attained by trained children only.

5.1 Introduction

Assessment of children's cognitive abilities is usually seen as a necessary step in monitoring children's progression in academic learning, often by means of conventional assessment instruments. Although using such conventional tests may have advantages, such as their predictive qualities, ease of administering, and clear outcomes, a number of theorists and practitioners have criticized the use of these instruments (e.g., Elliott, Grigorenko, & Resing, 2010; Fiorello et al., 2007). These criticisms include the limited information conventional tests provide regarding children's task solving processes (Richard & Zamani, 2003), and in relation to interventions targeted at enabling the use of cognitive potential more efficiently (Elliott et al., 2010; Elliott, Resing, & Beckmann, 2018). The current research aimed to investigate the usefulness of task solving process information in a dynamic visual-spatial complex figure task, and the effects of training on the processes used in solving this task, through the use of a rule-based, theory-driven scoring method for task solving processes.

Process-oriented dynamic testing

In contrast to conventional static testing, dynamic testing includes feedback or training in the test. By providing this feedback, dynamic testing aims to reveal both children's current level of cognitive functioning and their potential for learning (Elliott et al., 2010). Many theorists have underlined the importance of assessing process information while administering cognitive tasks, along with its products. Process information could potentially provide valuable information regarding interventions aimed at children learning more effectively (Greiff, Wüstenberg, & Avvisati, 2015; Greiff et al., 2013). Process-oriented dynamic testing is aimed at investigating children's task solving processes, and changes in the use of these processes as a result of training (Resing & Elliott, 2011). Several studies have investigated the use of, and changes in strategy use during dynamic testing (e.g., Hessels, Vanderlinden, & Rojas, 2011; Resing, Bakker, Pronk, & Elliott, 2016; Resing, Xenidou-Dervou, Steijn, & Elliott, 2012; Resing & Elliott, 2011). As Resing and colleagues (2016) indicated, information regarding children's strategy deployment can provide valuable information

on children's cognitive potential. The general aim of the current study was therefore to investigate the explanatory value of process measures in a dynamic complex figure drawing test.

The current study utilized a test-training-test dynamic testing design employing the graduated prompts approach (e.g., Resing & Elliott, 2011). The training consisted of a series of standardized graduated prompts that were hierarchically ordered, providing feedback to the child when he or she was not able to correctly answer an item independently. Due to the provision of prompts in a hierarchic fashion, this training procedure allows for investigating the different degrees of help children need when learning to solve new tasks (Resing & Elliott, 2011).

Previous research has suggested that graduated prompts training leads to improvements in test accuracy (Resing & Elliott, 2011; Stevenson, Heiser, & Resing, 2013), but also in task solving processes (Resing, Bakker, Pronk, & Elliott, 2017; Resing et al., 2012). Moreover, this training approach has been found to lead to transfer to different, but related, tasks within one and the same cognitive domain (Roth-van der Werf, Resing, & Slenders, 2002; Stad, Vogelaar, Veerbeek, & Resing, 2017). These studies focused on the effectiveness of graduated prompts on test scores and task solving processes within the domain of inductive reasoning. In this study, the effectiveness of graduated prompts was analyzed in relation to children's complex figure drawing skills, and it was investigated to what extent trained skills within complex figure drawing would transfer to the domain of inductive reasoning.

Complex figure drawing

Drawing complex figures is a frequently used method in (neuro)psychological assessment. Drawing performance can be used as a quick and sensitive method of assessing visuo-spatial abilities (La Femina, Senese, Grossi, & Venuti, 2009), such as visual perception and visual construction (Martens, Hurks, & Jolles, 2014). As La Femina and colleagues (2009) pointed out, visual construction requires a variety of cognitive visuo-spatial processes, including visuo-spatial perception and representation, working memory, and planning. Kirkwood, Weiler, Bernstein, Forbes, and Waber (2001) underlined the role of encoding the figure that is to be drawn. Task analytic models of copy drawing tasks divide the process of the task into several phases (Senese, De Lucia, & Conson, 2015). Firstly, the elements of

the figure as well as their relations have to be identified by dividing the figure into a number of smaller parts (e.g., Roncato, Sartori, Masterson, & Rumiati, 1987). Next, the drawing plan has to be prepared by determining strategies to guide the procedure of drawing. Then, the plan needs to be translated into execution through grapho-motor action. Finally, the model and the copy have to be compared, thereby monitoring the execution of the drawing (Senese et al., 2015). In the current study, a dynamic version of the Rey-Osterrieth Complex Figure Task was used. The training utilized in the current study employed graduated prompts techniques that were specifically tailored to the complex figure task, and aimed at analyzing the figure in parts (Akshoomoff & Stiles, 1995; Kirkwood et al., 2001; Resch, Keulers, Martens, van Heugten, & Hurks, 2018), thereby highlighting which elements of the complex figure could be grouped together.

Aims and research questions

In this study, a dynamic complex figures test was constructed and utilized, entailing a graduated prompts method training children on both the use of more general task solving skills and on skills specifically relevant to the task.

The study firstly focused on children's potential improvements in accuracy on the complex figure task, and the potential transfer of training to an inductive reasoning task. In relation to accuracy of drawing the complex figure task, it was expected that graduated prompts training would lead to more accuracy in drawing complex figure tasks than unguided practice opportunities only (Resing & Elliott, 2011; Resing, Touw, Veerbeek, & Elliott, 2017). With regard to potential transfer of trained skills to a task in a different domain, it was expected that the complex figure graduated prompts training would not lead to increased performance on a visual-spatial inductive reasoning task. This hypothesis was built on previous research, in which it was found that transfer of trained skills occurred only in relation to tasks that were highly related (Roth-van der Werf et al., 2002).

Secondly, the process of task solving was analyzed based on process measures, which were thought to be operationalizations of children's changes in their ability to represent the tasks to be solved (Halford, Wilson, & Phillips, 1998; Pretz, Naples, & Sternberg, 2003). We examined children's potential changes in the process of accurately drawing the complex figures as a result of repeated practice and training, by distinguishing subgroups

of children based on their initial task solving processes (i.e. grouping behavior). It was expected that the distribution of children across these subgroups would change from pre- to posttest and that, compared with control group children, trained children would show more grouping activities (using separate sequences of task elements), thereby progressing towards more sophisticated categories of grouping behavior than non-trained children (e.g., Halford et al., 1998; Resing et al., 2016).

5.2 Method

Participants

Participants in this study were 106 7-8-year old children ($M = 7.8$ years, $SD = 0.42$ years), 53 boys and 53 girls. The children attended the 2nd grade of 10 primary schools in the Netherlands. The schools were located in lower to middle class areas in the Netherlands, and were selected on the basis of their willingness to participate. The children's primary language spoken at school was Dutch. Prior to the testing procedure, informed consent was obtained from all parents. Six children did not attend all sessions of the study and were excluded from the analyses. The study outlines were approved by the local ethics committee.

Design and procedure

The study used a randomized blocking pretest/posttest design. Before dynamic testing started, the Raven's Standard Progressive Matrices test (Raven, Raven, & Court, 1998) was administered. Then, per school, children were blocked on the basis of their Raven scores. Pairs of children were randomly assigned to an experimental or an unguided control condition. During the pre- and posttest, both a visual-spatial inductive reasoning task and the Rey-Osterrieth complex figure were administered. During the sessions in between, children in the experimental condition were trained on a complex figure task twice, and children in the control group did not receive training but, within the same time frame, solved dot-completion tasks twice. Finally, a posttest was conducted, which again consisted of both series completion and the Rey-Osterrieth complex figure task. The pre- and posttest lasted about 45-60 minutes per child, and the training sessions between 10 and 25 minutes per child, depending on the number of prompts children needed and their

general task solving speed. All tests were administered individually by 18 well-trained psychology bachelor students in a quiet room in the child's school.

Materials

Raven's Standard Progressive Matrices. Raven's SPM was used as a measure of initial inductive reasoning ability. The test includes 60 items, and requires the children to detect which piece is missing out of a matrix (3x3) based on the elements and relationships within an item. The Raven test has been found to have an internal consistency coefficient of $\alpha = .83$ and a split-half coefficient of $r = .91$ (Raven et al., 1998).

Schematic picture series completion. The series completion task (Resing & Elliott, 2011; Resing, Touw et al., 2017) was used as a measure of children's inductive reasoning ability. Each item consisted of a line of 6 schematic puppets existing of discrete elements differing in gender, color, and pattern, and consisted of a head, 2 arms, 2 legs, and 3 body parts, where the child was required to construct the next (7th) puppet figure using the separate pieces. The 2x12 pre- and posttest items were constructed to be equivalent to each other, using identical rules for the changes in the puppet figures (Resing, Touw et al., 2017). In a recent study with the same, tangible materials internal consistencies for the pretest ($\alpha = .74$) and posttest ($\alpha = .78$), and a test-retest reliability ($r = .78$) were reported (Stad, Van Heijningen, Wiedl, & Resing, 2018).

Dynamic complex figure task. The Rey-Osterrieth Complex Figure test (Osterrieth, 1944; Rey, 1941) has been developed as a tool for measuring visual construction, visuo-spatial orientation, and memory (Senese et al., 2015). The figure consists of a complex line design, composed out of eighteen different features (Figure 1, left). The adapted, dynamic version of the complex figure task we constructed consisted of a pre- and posttest (the original figure was presented) and two training sessions, in which 2x2 newly constructed figures with different designs, but a similar number of features (see Figure 1, right part for an example) were used, and a posttest (the original figure again). During pre- and posttest children had to copy the presented figure with the model available on their desk, and then make an immediate recall drawing from memory. In the current study, we only used children's copy of the complex figure.

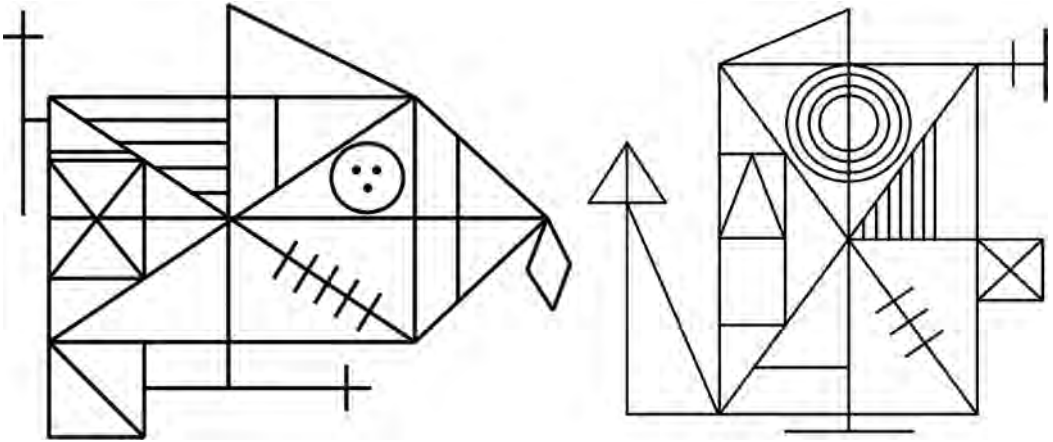


Figure 1. The Rey-Osterrieth complex figure (Osterrieth, 1944; Rey, 1941), and an example of the alternative training figures used during the complex figure graduated prompts training.

The graduated prompts training procedure consisted of two sessions including two newly constructed figures each. The hierarchical training started with general, metacognitive prompts, and progressed to more specific cognitive prompts, and, if necessary, modeling of the task solving process was offered. For each figure, children were asked to verbalize how they would draw the figure before they actually started drawing. Based on their verbal answers, children received further prompts, or were allowed to draw the corresponding part. The training procedure was based on dividing the figure into “basic structure”, and the “inside” and “outside” (Akshoomoff & Stiles, 1995; Kirkwood et al., 2001; Resch et al., 2018). To enable specific questioning, without providing too much a priori information to the child, the basic structure was split into two phases; drawing the rectangle and drawing the axes. The simplified training procedure is displayed in Appendix A.

Scoring and variables

Accuracy scores. During testing, the order of line drawing was recorded manually, and children’s accuracy score was based on a scoring method adapted by Taylor (1959). The 18 line-clusters of the figure were

appointed either 0, 0.5 or 1 point for accuracy of a specific line cluster and 0 or 1 point for correct placement of a cluster. The scores for accuracy and placement were added to create the total score, leading to a maximum score of 36 points for each rendering. The complex figures were scored by 4 well-trained psychology master students. Previous research has shown high inter-rater reliability for this scoring method (Loring, Martin, Meador, & Lee, 1990; Tupler, Welsh, Asare-Aboagye, & Dawson, 1995).

Task solving scores. To assess children's task solving processes, a Grouping of Answer Pieces (GAP) measure was defined, based on a breakdown of the sequence of task solving into smaller sub-sequences. To determine the parts of the figure that would be meaningful when grouped together, the figure was divided into its smallest parts, which were 56 lines. An initial division of drawing sequences of these lines into "basic structure", "inside", and "outside" (Kirkwood et al., 2001; Resch et al., 2018) was further refined and adapted into groups of line sequences, partially following the Boston Qualitative Scoring System (e.g., Akshoomoff & Stiles, 1995). This led to 24 different groups of lines, displayed in Appendix B. Each group was awarded 1 point if the lines within that group were drawn in immediate succession of each other. These 24 groups of lines were transformed into algorithms and programmed into Microsoft Excel as automated formulae, by which the number of groups the children drew could be calculated. The number of groups drawn was divided by the total amount of lines drawn, to create a corrected GAP-score.

GAP-categories. Based on the percentage of lines children drew within a single sequence of lines drawn, children were appointed to a GAP-solving category. The first category (Non-sequential) included the children that drew up to 20% of the lines of the figures within a sequence. The second category (Low mixed-sequential), required between 20-40% of the lines being drawn sequentially. If children showed sequential behavior in 40-60% of the lines drawn, they were appointed to the third category (Partial sequential). The children that used sequences in 60-80% of their lines, were appointed to the fourth category (High mixed-sequential), and those using the sequences for 80-100% of the lines were appointed to the highest category (Full sequential).

5.3 Results

Preliminary analyses were used to check a priori differences between the groups of children. Two separate one-way ANOVAs revealed that children in the different conditions did not differ in Raven scores ($p = .80$), nor in age ($p = .21$).

Training effects on accuracy of complex figure drawing and inductive reasoning

The effect of the graduated prompts training on children’s complex figure accuracy score was tested employing a repeated measures ANOVA, using children’s performance on the Complex Figure copy as the dependent variable, session (pretest/posttest) as the within-subjects factor, and condition (control/training) as the between-subjects factor (see Table 1 or Figure 2 for an overview of the means and standard deviations). Significant effects were found for session ($F(1, 104) = 28.63, p < .001, \eta_p^2 = .22$), condition ($F(1, 104) = 6.40, p = .013, \eta_p^2 = .06$), and session*condition ($F(1, 104) = 17.70, p < .001, \eta_p^2 = .15$). In line with our expectations, children that had received the graduated prompts training performed significantly better on the complex figure posttest than children who had not received training.

Similarly, the effect of the graduated prompts training on children’s series completion accuracy score was tested, using series completion task as the dependent variable, session (pretest/posttest) as the within-subjects factor, and condition (control/training) as the between-subjects factor. No significant effects were found for session ($p = .94$), condition ($p = .29$), or session*condition ($p = .57$). As expected, the complex figure graduated prompts training did not lead to transfer in series completion performance.

Table 1. Means and standard deviations for performance on the series completion task and the Rey-Osterrieth Complex Figure task for the different conditions.

	Dynamic testing N=54		Unguided control N=52	
	Pre <i>M</i> (SD)	Post <i>M</i> (SD)	Pre <i>M</i> (SD)	Post <i>M</i> (SD)
Complex figure	21.25 (6.63)	27.28 (6.12)	20.75 (7.25)	21.47 (8.60)
Series completion	5.09 (2.33)	5.24 (2.87)	4.77 (2.36)	4.65 (2.42)

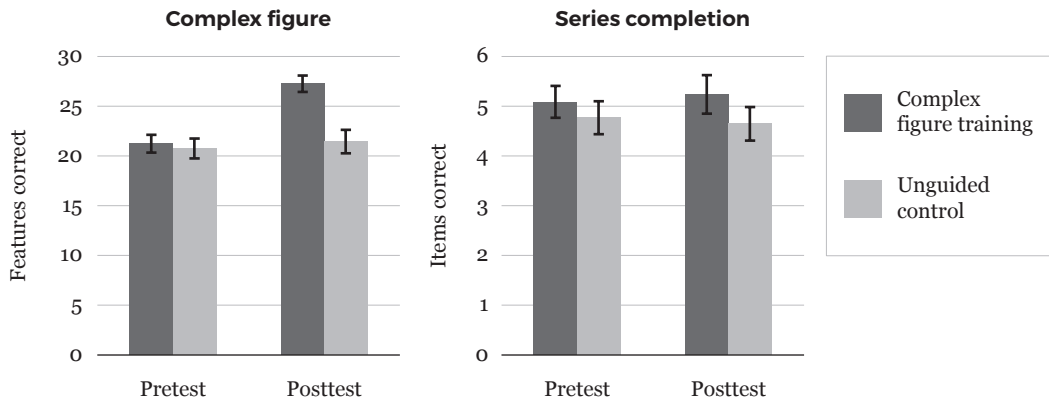


Figure 2. Graphic representations of children's pre- and posttest score on Rey-Osterrieth Complex Figure and Series completion (per condition) including standard error.

Effects of graduated prompts training on grouping behavior

Next, children's changes in grouping behavior (GAP) on the complex figure task were analyzed. The relationship between condition and children's allocations to complex figure GAP-categories was examined by means of a χ -square test (see Table 2 for the outcomes). On the pretest, no significant association was found between condition and complex figure GAP ($\chi^2_{\text{pretest}} (n=106) = 2.09, p = .554, 25.0\%$ of the cells have expected count less than 5). On the posttest, as expected, a significant association was revealed between condition and complex figure GAP ($\chi^2_{\text{posttest}} (n=106) = 63.52, p < .001, 0\%$ of the cells have expected count less than 5). Children's solving behavior in terms of GAP significantly progressed from pre- to posttest as a consequence of the training. Children in the control condition showed only a small shift from pre- to posttest, whereas trained children no longer used the lowest 2 categories on the posttest, and only trained children attained the highest GAP-category.

Table 2. Results for the crosstabs analyses for grouping of answer pieces.

		1. Non- sequential	2. Low mixed	3. Partial sequential	4. High mixed	5. Full sequential	Total
<i>Dynamic test condition</i>							
Pretest	Frequency	3	15	30	6	0	54
	Percentage	5.6	27.8	55.6	11.1	0.0	100
Posttest	Frequency	0	0	4	34	16	54
	Percentage	0.0	0.0	7.4	63.0	29.6	100
<i>Unguided control condition</i>							
Pretest	Frequency	2	21	23	6	0	52
	Percentage	3.8	40.4	44.2	11.5	0.0	100
Posttest	Frequency	0	15	28	9	0	52
	Percentage	0.0	28.8	53.8	17.3	0.0	100

5.4 Discussion

In this study, a dynamic complex figures test was constructed and utilized, entailing a graduated prompts training, which provided children with more general task solving prompts and scaffolds that were specifically relevant to the task. The construction of the dynamic complex figure task included the use of a rule-based theory driven scoring method for task solving processes.

Firstly, the effects of the graduated prompts training were investigated. This training procedure was developed to support children’s organization of drawing a complex figure task, and was thought to support children in representing the figure more efficiently (Halford et al., 1998). The specific prompts of the training procedure were partially based on the principle of organization, as used in prior research (e.g., Akshoomoff & Stiles, 1995; Kirkwood et al., 2001; Resch et al., 2018). Our results showed that trained children significantly improved in efficiency when drawing the complex figure, whereas untrained children did not. This was in line with previous research in other domains where graduated prompts training was found to be an effective method for dynamic testing, such as analogical reasoning

(e.g., Stevenson et al., 2013), and series completion (e.g., Stad et al., 2017). In line with previous research indicating that transfer was only found for tasks that were highly related (Roth-van der Werf et al., 2002), the effects of the graduated prompts training remained limited to the immediate domain children were trained on, and did not lead to transfer of the learned skills into the more general domain of visual-spatial inductive reasoning.

Secondly, the effects of the graduated prompts training on children's solving behavior were investigated, through the use of a rule-based, theory-driven scoring method. This method was expected to provide information on the way in which children grouped the lines they had to draw, thereby uncovering, at least partially, children's organization of their mental representation of the figure (Halford et al., 1998, Pretz et al., 2003). Through using this process-oriented dynamic testing procedure, we were able to detect changes in the task solving processes children used (Resing & Elliott, 2011). Although a small shift in solving behavior was visible for untrained children, children trained in drawing a complex figure showed to a larger extent that they organized their drawing, thereby efficiently attaining an advanced line-grouping level on the posttest. These findings indicate that they were able to change their task solving behavior, possibly as a result of more efficient task representations. These findings support Halford and colleagues' (1998) statement that didactic help is necessary for children to develop strategies for more efficiently storing information into their mental representations. After training, most children organized the majority of the lines in a configural way. The training was not only beneficial for children who already used some form of grouping on the pretest, but also for those who initially showed no use of grouping behavior. Trained children appeared to be more able to represent the figure in a configural way, thereby reducing the load of the representation of the figure (Halford et al., 1998).

The current findings have to be interpreted in the light of some limitations of this research. The relatively low number of participants and the limited age-range of the participating children, imply that our results cannot be generalized to different age-ranges without further, detailed research. In addition, although the process scores were obtained through rule-based automated scoring using a computerized algorithm, the data used for this scoring procedure were recorded manually, which may have

negatively affected their accuracy. Future research could focus on devising new techniques to record data in complex (drawing) tasks in a computerized way, as well as provide automated scoring and possibly support in interpreting the obtained data.

Our results provide support for the notion that process assessment is applicable to new tasks, as long as the underlying cognitive processes involved in successfully solving the task are known. Before rule-based, theory-driven automated scoring can be applied to tasks, however, the processes assessed within solving the task have to be analyzed in detail (Csapó & Funke, 2017). The assessment of task solving processes in education might, in principle, open possibilities to guide aimed intervention, by providing information on which task solving processes are used efficiently by children, and which are not. Future research might incorporate tailored intervention, based on the child's task solving processes. Through the use of computers, bespoke training procedures can be used as a reaction to children's processes during task solving. Future research could also focus on individual differences in the use of the strategies taught during the training phase. This might provide valuable information on children's response to intervention in terms of the cognitive processes they were able to apply (Resing et al., 2012).

APPENDIX A.

Graduated prompts training procedure for the complex figure (simplified).

Phase	Prompt
<i>Base of the figure (rectangle)</i>	<i>Before you start drawing I would like to know where you will start.</i> 1. What would be useful to start with, so that the rest of the figure can be attached to it? 2. What is the base of the figure? One simple big shape the other parts are attached to? 3. The figure contains a large rectangle that forms the base of the figure, can you find it? 4. Modelling, pointing out the large rectangle.
<i>Division of the figure (axes)</i>	<i>Now that we have drawn the base, how can we continue?</i> 1. How can we expand on the base, so that we have a structure to further build on? 2. What else belongs to the basic structure? Are there simple lines that divide the figure? 3. There are some long lines on the figure, that divide it. Can you find them? 4. Modelling, pointing out the axes.
<i>Division of the elements (in/out)</i>	<i>Now that we have drawn the basic structure, how can we structure the rest?</i> 1. Look at what we've already drawn. Can that help in dividing the elements? 2. Look at the base of the figure. Can we use the rectangle to divide the elements? 3. The large rectangle makes the base. Everything else is attached to it. Could you use inside or outside of the rectangle to divide the elements? 4. Modelling, pointing out the elements inside and outside of the rectangle.

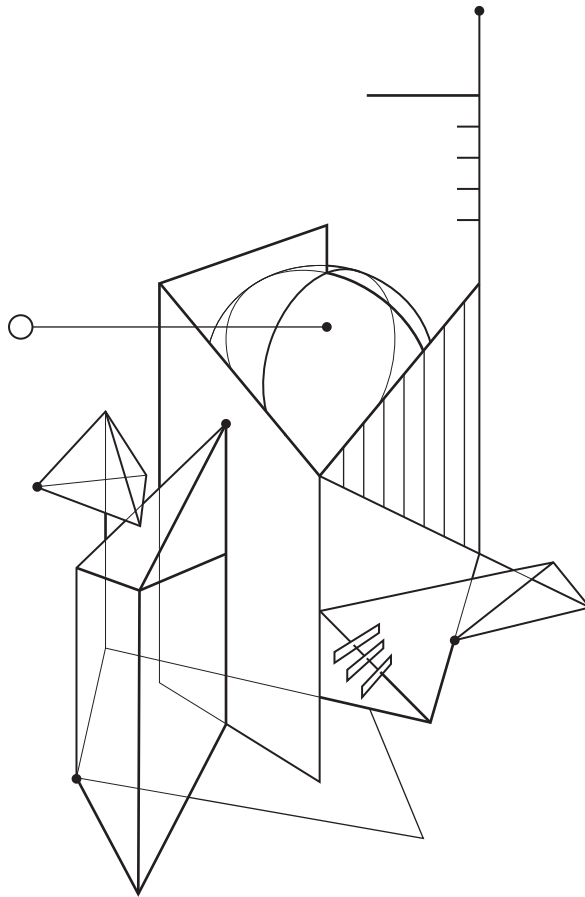
APPENDIX B.

The groups of line that were discerned for defining the pre- and post- GAP-measure (Akshoomoff & Stiles, 1995; Kirkwood et al., 2001; Resch et al., 2018).

Basic structure		
1. Rectangle, upper side		
2. Rectangle, left side		
3. Rectangle, bottom side		
4. Rectangle, right side		
5. Full rectangle		
6. Diagonal bottom left		
7. Diagonal bottom right		
8. Horizontal axis		
9. Vertical axis		
10. Diagonals		
11. Axis		
12. Web		
13. Complete basic structure		
Outside		
14. Hypotenuse		
15. Triangle with diamond and lines		
16. Cross bottom		
17. Square bottom		
18. Cross + square		
19. Cross left		
Inside		
20. Hash marks		
21. Horizontal lines		
22. Small rectangle		
23. Diagonals in small rectangle		
24. Small rectangle + diagonals		

CHAPTER 6

General discussion



The main goal of this thesis was to investigate the information that can be obtained from process-oriented dynamic testing using a rule-based, theory-driven scoring method. More specifically, the specific characteristics of Grouping of Answer Pieces (GAP), a method of computer automated process assessment based on children's sequences of solving activities when answering cognitive problems, were object of study.

In this chapter, a summary of the findings from the studies described in this thesis was provided, and the theoretical considerations and practical implications of these findings were discussed. Then, the limitations in our studies were reviewed, and finally recommendations for future research were provided.

Summary of findings

The main aim of the first study presented in this thesis (*Chapter 2*) was evaluating a new measure for analyzing task solving processes in series completion tasks in 8-year-old children. The measure, *Grouping of Answer Pieces* (GAP), was thought to provide information on children's problem representation and restructuring. Information regarding children's task solving processes was collected, utilizing an electronic tangible user interface was. The GAP measure was found to be a moderate predictor of accuracy in solving the series completion tasks, as were verbalized strategy use and time measures. Analyses at the level of individual items revealed that, depending on item difficulty, the GAP measure was the only process measure that was moderately related to successfully solving the task. Furthermore, the GAP measure was not significantly related to verbalized strategy use and time measures, indicating that it measured unique variance. In addition, more stable grouping behavior was related to better series completion solving ability. In contrast, for verbalized strategy use, more variability was related to better task outcomes. Taken together, these findings support the notion that these measures provided information on different aspects of the task solving process.

Chapter 3 focused on the effects of using or not using a pretest in a dynamic series completion task in 7-to-8 year old children. Half of the children received a pretest, a graduated prompts training and a posttest, the

other half only received the graduated prompts training and the posttest. No differences were found between children that had received a pretest and children that had not, in terms of series completion solving ability on the posttest, the processes used in solving the posttest items in terms of the GAP measure, verbalized strategy use, and planning time, or the number of hints they needed during training. Further analyses showed that process measures, more specifically the GAP measure and verbalized strategy use, predicted item success. The lack of difference between children that had and had not received a pretest, indicated that the decision of whether or not to use a pretest could be based on the diagnostic questions to be answered and the testing situation, rather than on psychometric considerations.

Chapter 4 investigated 7-to-8 year old children's task solving processes and changes in these processes as a result of a graduated prompts training in a dynamic series completion task. Half of the children received a graduated prompts training, the other half received repeated practice only. Children's task solving processes were measured in terms of the GAP measure, verbalized strategy use, and completion time. Trained children showed more progress in their series completion solving ability from pre- to posttest than untrained children did. In relation to task solving processes, the effects of training were most visible in children's verbalized strategy use, which became more advanced as a result of training. With regard to the GAP measure, trained children did not show a more advanced level of use than untrained children, but they did show different distributions over the different GAP categories. There were no differences in completion time of children who were trained or only received repeated practice. Although the process measures were found to be related to performance on the dynamic series completion task, as well as to math and reading comprehension performance, the amount of help children received during training had most predictive value for school performance.

In *Chapter 5* the use of process-oriented dynamic testing was investigated within a different domain, again using a pretest/training/posttest design, where half of the children received a training, and the other half received repeated practice only. In this chapter, a visual-spatial complex figure task was used to investigate the effects of training on the processes on and outcomes of this task with 7-to-8 year old children. A graduated prompts

training procedure was developed, aimed at improving children's organization of the figure. Training led to more progression from pre- to posttest in terms of children's complex figure drawing performance, as well as use of more advanced grouping behavior. These findings indicated a shift in processes used to complete the complex figure drawing task, which seemed to have been brought about by the graduated prompts training. Only trained children were able to attain the highest level of organization of the figure.

Theoretical and practical considerations

The studies of this thesis were focused on process-oriented dynamic testing (e.g., Resing, Bakker, Pronk, & Elliott, 2017; Resing, Xenidou-Dervou, Steijn, & Elliott, 2012; Resing & Elliott, 2011). In-depth dynamic assessment was thought to unveil more information regarding children's instructional needs. In this thesis, process-oriented measurement was found to provide valuable information beyond outcome scores, including the effects of training on a child's task solving processes.

Dynamic testing

The majority of the studies in this thesis (*Chapter 3, 4, and 5*) made use of dynamic testing, which was executed using a graduated prompts training procedure (e.g., Resing, 2013). Dynamic testing was found to lead to significant increases in children's solving performance in both a series completion task (*Chapter 4*), and a complex figure drawing task (*Chapter 5*). These findings were in line with previous research that successfully used the graduated prompts approach in dynamic testing of, among others, series completion ability (e.g., Resing et al., 2012; Stad, Vogelaar, Veerbeek, & Resing, 2017), analogical reasoning (e.g., Stevenson, Heiser, & Resing, 2013), and language learning (e.g., Camilleri & Botting, 2013; Hasson, Dodd, & Botting, 2012). Taken together, these findings seem to support the notion that the graduated prompts training method can lead to improved task performance, irrespective of the task domain. Furthermore, in line with previous research (Caffrey, Fuchs, & Fuchs, 2008; Stevenson, Bergwerff, Heiser, & Resing, 2014; Swanson & Howard, 2005), dynamic testing measures in series completion were found to provide moderately more explained variance than static testing measures

(*Chapter 4*). Particularly the number of hints children needed during training provided a better prediction of children's academic performance than static measures, in line with previous research (e.g., Stevenson et al., 2014).

Despite the promising features of dynamic testing, a question has been raised concerning the influence of using a pretest/posttest design on the validity and reliability of the test scores (Kim & Willson, 2010; Klauer, 1993; Sijtsma, 1993; Sternberg & Grigorenko, 2002). In *Chapter 3*, the test performances of children receiving a full dynamic test (pretest/training/posttest) and children receiving training and posttest only, were compared to address this issue. No differences were found in children's outcomes on series completion tasks, nor in the processes children employed to solve the task. This led us to conclude that for the graduated prompts dynamic series completion test, the use of a pretest did not lead to changes to the reliability or validity of the test. As a result, when working with dynamic testing in practice, when deciding whether or not to use a pretest, the question to be answered by that test might be used in the first place rather than psychometric differences in the construct measured by the test.

Process assessment

Several process measures were utilized to investigate children's task solving processes during dynamic testing in series completion. Children's verbal explanations of their task solving processes were moderately related to accuracy on the series completion task (*Chapter 2, 3, and 4*), and appeared to be most in line with the characteristics of strategy use as described by Siegler (1996, 2007), who defined strategy use as variable, both between and within tasks. Variability in verbalized strategy use was related to greater accuracy in the series completion task (*Chapter 2*), indicating that, in line with the theory of Hunt (1980), rather than being connected to one particular strategy, successful task solving behavior is connected to the ability to adaptively alternate between strategies if the tasks requires this. Children's verbalized strategy use progressed towards more sophisticated explanations as a result of training (*Chapter 4*). Apparently, the learning situation did not only lead to improved performance, but also more advanced strategy use, which was in line with the interactive relationship between strategy use and learning, as described by Alibali, Phillips, and Fischer (2009).

Compared with the other available process measures, completion time appeared to have the weakest relationship with accuracy on the series completion task (*Chapter 2 and 4*), and was not influenced differently by training than by repeated practice (*Chapter 4*). This may have been a result of the level of difficulty of the tasks, which in the past has been shown to influence the relationship between completion time and accuracy (Dodonova & Dodonov, 2013; Goldhammer et al., 2014). The time taken for planning correlated only moderately with accuracy (*Chapter 2*), whereas accuracy was related more strongly to the verbalization and GAP measures.

Grouping of answer pieces (GAP). The grouping of answer pieces in children's task solutions was thought to be another process measure, being indicative of the way in which children represented a problem, and of strategies that were employed to more efficiently store this information, by grouping related information together into "chunks" (Halford, Wilson, & Phillips, 1998; Pretz, Naples, & Sternberg, 2003). In line with the theory of Sternberg (1985), who stated that problem representation is a meta-component of problem solving, more stable scores on the GAP measure, as an indicator of children's problem representation, were related to better task outcomes. The GAP measure was consistently found to be a moderate predictor of item success (*Chapter 2, 3, and 4*). In line with the literature, the task representation as measured with the GAP measure was closely related to successful task solving (Hunt, 1980; Pretz et al., 2003; Robertson, 2001).

Restrictions of children's task representation have been related to limitations of their working memory capacity. The use of more effective strategies for storing information has been argued to lead to a more efficient mental representation of the task (e.g., Andrews & Halford, 2002; Halford et al., 1998). Halford and colleagues (1998) concluded that if children do not have these strategies available, didactic support is a prerequisite to help children more efficiently represent the information necessary for solving the task. In line with this research, children's representations of the complex figure task as operationalized with the GAP measure, improved as a result of a targeted graduated prompts training procedure, but not as a result of repeated practice (*Chapter 5*). The series completion study revealed a more differentiated picture regarding the different process measures utilized. Trained children profited from the dynamic training given to them and showed more

progress in their series completion solving ability from pre- to posttest than their control group peers. Under influence of training, their verbalization measures progressed towards a more advanced level. The GAP measures of both trained and untrained children progressed as well. Both dynamic tests used in this thesis were very different from each other, including the training scripts provided to the children. For the moment, we have to conclude that task content, task difficulty, and variation in training procedures influence the role of GAP measures in process-oriented dynamic testing.

The overall picture that emerged from the findings support the notion that each process measure reveals a different part of the actual solving steps children showed when solving complex cognitive tasks. Grouping behavior, as measured by the GAP measure, certainly represents one aspect of the problem representation as described by Sternberg (1985), and can be thought to be a stable metacomponent, which serves to guide the task solving process (*Chapter 2*). Some children, however, might need a didactic intervention to efficiently use the problem representation (e.g., Halford et al., 1998).

Assessment of children's cognitive abilities. Throughout this thesis, a new measure was used, which would have been very difficult to obtain without the use of computers. The GAP measure was developed from an algorithm which detected the grouping of pieces in the overall sequence of placement or drawing of the answer. The basic principle was developed first for a series completion task, based on extensive task analysis. Later, the same basic principle was redeveloped for a complex figure task, based on previous research of the processes involved in solving complex figure drawing tasks. The use of constructed response items has been advocated in the past, as they provide more and more detailed information on children's task solving abilities and processes (Stevenson, Hickendorff, Resing, Heiser, & de Boeck, 2013; Yang, Buckendahl, Juszkievicz, & Bhola, 2002). Through combining them with computerized systems, the nature of scoring the tasks becomes less labor-intensive, and, as was found in this thesis, information on children's task solving processes can be obtained relatively easily.

Tangible User Interfaces. Through the use of tangible user interfaces (TUIs), the benefits of computers were combined with the benefits of tangible materials for children's development, and to elicit authentic behavior from the children (Verhaegh, Resing, Jacobs, & Fontijn, 2009). Using such

computerized systems enabled children to work independently and receive feedback and training from the console (Resing et al., 2012; Verhaegh, Fontijn, Aarts, & Resing, 2013). It also provided the possibility of log-file analysis as a source of information on children's task solving processes. Combined with a rule-based, theory-driven automated scoring system, it could provide educators and educational psychologists with easy to obtain, in-depth information on children's task solving behavior, with a minimum of time and effort that needs to be invested by an examiner. This might enable process-oriented dynamic testing in the classroom, without any severe time investment of educators or psychologists, an often-voiced objection to the widespread use of dynamic testing (Sternberg & Grigorenko, 2002).

Analysis of process measures. For the analysis of the processes children use to solve certain tasks, Siegler (1987) pointed out that averaging data over multiple items could lead researchers to overlook aspects of the task solving processes. Investigating these task solving processes on the level of individual items (*Chapter 2 and 3*) was expected to provide valuable information regarding the potential factors contributing to children's success or failure when tested. The results underlined the importance of task characteristics in interpreting children's task solving processes. Previous research showed that the interpretation of completion time is dependent on task characteristics such as item difficulty (e.g., Dodonova & Dodonov, 2013; Goldhammer et al., 2014; Scherer, Greiff, & Hautamäki, 2015), and in line with this research, item characteristics were found to be an important explanatory factor of item success. In analyzing these complex and non-linear relationships, decision tree analyses provided a useful tool (Ritschard, 2014), and provided additional insights when combined with traditional, linear analyses.

Limitations and future research

Our findings should be viewed in the light of a number of limitations of the studies described in this thesis. Firstly, participants in all studies were primary school children of a single age range (7-to-8 years of age). More research might be needed to apply our findings to a more diverse population in terms of age, but also for the application in special education contexts.

The assessment of children in special education requires specific expertise, and it cannot be readily assumed that the processes that are used to solve the tasks are the same for children in special education contexts (Hessels, Vanderlinden, & Rojas, 2011).

Additionally, despite using TUIs to obtain non-obtrusive measurement of children's problem solving processes, children were also required to provide verbal explanations of their task solution. Requiring children to verbally explain their answers, may have influenced their task solving processes (Kirk & Ashcraft, 2001; Tenison, Fincham, & Anderson, 2014). To control for this, future research might employ a control group that is not required to verbally explain their answer, to provide more insight into the influence of the requirement of verbal explanations on children's task solving processes.

Greiff, Wüstenberg, and Avisati (2015) discerned a number of factors that might prevent widespread use of log-file analysis, such as the complicated relationships between data, difficulty to determine the exact meaning of data patterns and their implications, the lack of technical expertise, and the fact the theory-driven research has never gained momentum. The complicated relationships between data, such as interactive relationships with task characteristics (e.g., Goldhammer et al., 2014), or curvilinear relationships between process measures and accuracy (e.g., Greiff, Niepel, Scherer, & Martin, 2016) could distort our understanding of the processes children use to solve tasks. To overcome this, future research might not only focus on linear relationship, but use more diverse methods to support our understanding of task solving processes, such as Rasch modeling (Stevenson, Hickendorff, et al., 2013), SEM modeling (Greiff et al., 2016), or decision tree analysis using C(A)RT and CHAID methods (McArdle, 2014; Ritschard, 2014). Additionally, Siegler (1987), recommended to look at the contributions to item success and be mindful of the dangers of averaging data over multiple items.

To extract exact meaning from the process measures, and allow for their application in recommendations towards educational interventions, future research might investigate the connections between task solving processes children use, and interventions that can be used to remediate processes that are not used efficiently. Through combining the processes within a specific domain of testing, with a specific intervention that targets

the most relevant processes, assessment and intervention can be connected from an empirical basis, and provide a solid foundation for application in education. Future research might focus on the development and use of fully computerized dynamic tests, for instance through computerized versions of the Rey-Osterrieth complex figure, which would allow for automatized scoring of both accuracy and process measures. Through working with technology independently, children's learning potential and task solving processes can be uncovered, whilst decreasing the load for teachers. Using the technological opportunities that the 21st century has brought us might prove challenging, but may also provide valuable tools for the assessment and development of individual potential.

References

A

- Akshoomoff, N. A., & Stiles, J. (1995). Developmental trends in visuospatial analysis and planning: I. Copying a complex figure. *Neuropsychology, 9*(3), 364–377.
- 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.
- Andrews, G., & Halford, G. S. (2002). A cognitive complexity metric applied to cognitive development. *Cognitive Psychology, 45*(2), 153–219.

B

- Benson, N., Hulac, D. M., & Bernstein, J. D. (2013). An independent confirmatory factor analysis of the Wechsler intelligence scale for children-fourth edition (WISC-IV) integrated: what do the process approach subtests measure? *Psychological Assessment, 25*(3), 692–705.

C

- Caffrey, E., Fuchs, D., & Fuchs, L. S. (2008). The predictive validity of dynamic assessment: A review. *The Journal of Special Education, 41*(4), 254–270.
- Camilleri, B., & Botting, N. (2013). Beyond static assessment of children's receptive vocabulary: The dynamic assessment of word learning (DAWL). *International Journal of Language and Communication Disorders, 48*(5), 565–581.
- Campione, J. C. (1989). Assisted assessment: A taxonomy of approaches and an outline of strengths and weaknesses. *Journal of Learning Disabilities, 22*(3), 151–165.
- Campione, J. C., & Brown, A. L. (1978). Toward a theory of intelligence: Contributions from research with retarded children. *Intelligence, 2*(3), 279–304.

- 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.
- Campione, J. C., Brown, A. L., Ferrara, R. A., Jones, R. S., & Steinberg, E. (1985). Breakdowns in flexible use of information: Intelligence-related differences in transfer following equivalent learning performance. *Intelligence*, 9(4), 297–315.
- Cockcroft, K., Alloway, T., Copello, E., & Milligan, R. (2015). A cross-cultural comparison between South African and British students on the Wechsler Adult Intelligence Scales Third Edition (WAIS-III). *Frontiers in Psychology*, 6, 1–11.
- Colman, A. M. (2006). *Dictionary of Psychology*. New York: Oxford University Press.
- Csapó, B., & Funke, J. (2017). The development and assessment of problem solving in 21st-century schools. In B. Csapó & J. Funke (Eds.), *The nature of problem solving: Using research to inspire 21st century learning* (pp. 19–32). Paris, France: OECD Publishing.

D

- De Koning, E., & Hamers, J. H. M. (1999). Teaching inductive reasoning: Theoretical background and educational implications. In J. H. M. Hamers, J. E. H. van Luit, & B. Csapó (Eds.), *Teaching and Learning Thinking Skills* (pp. 157–188). Lisse, The Netherlands: Swets & Zeitlinger B.V.
- Dienes, Z. (1964). *Building up mathematics* (2nd Ed.). London: Hutchinson Educational.
- 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.
- Dourish, P. (2004). What we talk about when we talk about context. *Personal and Ubiquitous Computing*, 8(1), 19–30.

E

- Elliott, J. G. (2000). The psychological assessment of children with learning difficulties. *British Journal of Special Education*, 27(2), 59–66.

- 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.
- 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., & Beckmann, J. F. (2018). Dynamic assessment: A case of unfulfilled potential? *Educational Review*, 70(1), 7–17.
- 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.
- Ericsson, K. A., & Simon, H. A. (1980). Verbal reports as data. *Psychological Review*, 87, 215–251.

F

- Fabio, R. A. (2005). Dynamic assessment of intelligence is a better reply to adaptive behavior and cognitive plasticity. *The Journal of General Psychology*, 132(1), 41–66.
- Fagan, J. F., & Holland, C. R. (2007). Racial equality in intelligence: Predictions from a theory of intelligence as processing. *Intelligence*, 35(4), 319–334.
- 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.
- 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.
- Feuerstein, R., Feuerstein, R. S., Falik, L. H., & Rand, Y. (2002). *The dynamic assessment of cognitive modifiability*. Jerusalem, Israel: The ICELP Press.
- Fiorello, C. A., Hale, J. B., Holdnack, J. A., Kavanagh, J. A., Terrell, J., & Long, L. (2007). Interpreting intelligence test results for children with disabilities: Is global intelligence relevant? *Applied Neuropsychology*, 14(1), 2–12.
- Fireman, G. (1996). Developing a plan for solving a problem: A representational shift. *Cognitive Development*, 11(1), 107–122.

G

- Goldhammer, F., Naumann, J., Rölke, H., Stelter, A., & Tóth, K. (2017). Relating product data to process data from computer-based competency assessment. In D. Leutner, J. Fleischer, J. Grünkorn, & E. Klieme (Eds.), *Competence assessment in education* (pp. 407–425). Cham: Springer.
- 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.
- Goswami, U. (2008). *Cognitive development: The learning brain*. Hove, UK: Psychology Press.
- Greiff, S., Niepel, C., Scherer, R., & Martin, R. (2016). Understanding students' performance in a computer-based assessment of complex problem solving: An analysis of behavioral data from computer-generated log files. *Computers in Human Behavior*, 61, 36–46.
- Greiff, S., Wüstenberg, S., & Avvisati, F. (2015). Computer-generated log-file analyses as a window into students' minds? A showcase study based on the PISA 2012 assessment of problem solving. *Computers and Education*, 91, 92–105.
- Greiff, S., Wüstenberg, S., Molnár, G., Fischer, A., Funke, J., & Csapó, B. (2013). Complex problem solving in educational contexts – Something beyond g: Concept, assessment, measurement invariance, and construct validity. *Journal of Educational Psychology*, 105, 364–379.
- Grigorenko, E. L. (2008). Dynamic assessment and response to intervention: Two sides of one coin. *Journal of Learning Disabilities*, 42(2), 111–132.
- Grigorenko, E. L., & Sternberg, R. J. (1998). Dynamic testing. *Psychological Bulletin*, 124(1), 75–111.

H

- Halford, G. S., Wilson, W. H., & Phillips, S. (1998). Processing capacity defined by relational complexity: implications for comparative, developmental, and cognitive psychology. *The Behavioral and Brain Sciences*, 21(6), 803–864.
- Halford, G. S., Wilson, W. H., & Phillips, S. (2010). Relational knowledge: The foundation of higher cognition. *Trends in Cognitive Sciences*, 14(11), 497–505.

- Hasson, N., Dodd, B., & Botting, N. (2012). Dynamic assessment of sentence structure (DASS): Design and evaluation of a novel procedure for the assessment of syntax in children with language impairments. *International Journal of Language and Communication Disorders*, 47(3), 285–299.
- Haywood, C. H., & Lidz, C. S. (2007). *Dynamic assessment in practice: Clinical and educational applications*. New York: Cambridge University Press.
- Henning, J. R., Verhaegh, J., & Resing, W. C. M. (2010). Creating an individualised learning situation using scaffolding in a tangible electronic series completion task. *Educational & Child Psychology*, 28(2), 85–100.
- 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.
- 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.
- Holzman, T. G., Pellegrino, J. W., & Glaser, R. (1983). Cognitive variables in series completion. *Journal of Educational Psychology*, 75(4), 603–618.
- Hunt, E. (1980). Intelligence as an information-processing concept. *British Journal of Psychology*, 71, 449–474.

J

- Janssen, J., Hop, M., & Wouda, J. (2015). *Wetenschappelijke verantwoording Rekenen-Wiskunde 3.0 voor groep 4* [Scientific report of arithmetic-mathematics 3.0 for grade 2]. Arnhem, The Netherlands: Cito BV.
- Jolink, A., Tomesen, M., Hilte, M., Weekers, A., & Engelen, R. (2015). *Wetenschappelijke verantwoording Begrijpend lezen 3.0 voor groep 4* [Scientific report of Reading Comprehension 3.0 for grade 2]. Arnhem, The Netherlands: Cito BV.

K

- Karpov, Y. V., & Tzuriel, D. (2009). Dynamic assessment: Progress, problems, and prospects. *Journal of Cognitive Education and Psychology*, 8(1), 228–237.

- 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* [Actuality and quality of norms: Method of determining norms of a student tracking system]. Arnhem, The Netherlands: Cito.
- 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).
- Kim, E. S., & Willson, V. L. (2010). Evaluating pretest effects in pre-post studies. *Educational and Psychological Measurement*, 70(5), 744–759.
- Kirk, E. P., & Ashcraft, M. H. (2001). Telling stories: the perils and promise of using verbal reports to study math strategies. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, 27(1), 157–175.
- Kirkwood, M. W., Weiler, M. D., Bernstein, J. H., Forbes, P. W., & Waber, D. P. (2001). Sources of poor performance on the Rey-Osterrieth Complex Figure Test among children with learning difficulties: a dynamic assessment approach. *The Clinical Neuropsychologist*, 15(3), 345–356.
- Klauer, K. J. (1993). Learning potential testing: The effect of retesting. In J. H. M. Hamers, K. Sijtsma, & A. J. J. M. Ruijsenaars (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.
- 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.
- Kossowska, M., & Necka, E. (1994). Do it your own way: Cognitive strategies, intelligence, and personality. *Personality and Individual Differences*, 16(1), 33–46.
- Kozulin, A. (2011). Learning potential and cognitive modifiability. *Assessment in Education: Principles, Policy and Practice*, 18(2), 169–181.
- Kuo, B.-C., Chen, C.-H., Yang, C.-W., & Mok, M. M. C. (2016). Cognitive diagnostic models for tests with multiple-choice and constructed-response items. *Educational Psychology*, 36(6), 1115–1133.

L

- La Femina, F., Senese, V. P., Grossi, D., & Venuti, P. (2009). A battery for the assessment of visuo-spatial abilities involved in drawing tasks. *Clinical Neuropsychologist*, 23(4), 691–714.
- Lidz, C. S. (2014). Leaning toward a consensus about dynamic assessment. Can we? Do we want to? *Journal of Cognitive Education and Psychology*, 13(3), 292–307.
- Loring, D. W., Martin, R. C., Meador, K. J., & Lee, G. P. (1990). Psychometric construction of the Rey-Osterrieth complex figure: Methodological considerations and interrater reliability. *Archives of Clinical Neuropsychology*, 5(1), 1–14.

M

- Manches, A., O'Malley, C., & Benford, S. (2009). The role of physical representations in solving number problems: a comparison of young children's use of physical and virtual materials. *Computers & Education*, 54 (3), 622–640.
- Marshall, P. (2007). Do tangible interfaces enhance learning? *Proceedings of the 1st International Conference on Tangible and Embedded Interaction* (pp. 163–170). New York, NY: ACM Press.
- Martens, R., Hurks, P. P. M., & Jolles, J. (2014). Organizational strategy use in children aged 5–7: Standardization and validity of the Rey Complex Figure Organizational Strategy Score (RCF-OSS). *The Clinical Neuropsychologist*, 28(6), 954–73.
- 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.
- Miller, G. A. (1994). The magical number seven, plus or minus two: some limits on our capacity for processing information. *Psychological Review*, 101(2), 343–352.
- 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.

N

- Neisser, U., Boodoo, G., Bouchard, T. J., Boykin, A. W., Ceci, S. J., Loehlin, J. C., & Sternberg, R. J. (1996). Intelligence: Knowns and unknowns. *American Psychologist*, 51(2), 77–101.
- Newell, A., & Simon, H. A. (1972). *Human problem solving*. Englewood Cliffs, NJ: Prentice-Hall.

O

- 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.
- Osterrieth, P. A. (1944). Le test de copie d'une figure complexe: Contribution à l'étude de la perception et de la mémoire [Copying a complex figure: Contributions to the study of perception and memory]. *Archives de Psychologie*, 30, 206–356.

P

- Perret, P. (2015). Children's inductive reasoning: Developmental and educational perspectives. *Journal of Cognitive Education and Psychology*, 14(3), 389–408.
- Piaget, J. (1976). *To understand is to invent. The future of education*. New York: Penguin Books.
- 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.
- Price, S., Jewitt, C., & Crescenzi, L. (2015). The role of iPads in pre-school children's mark making development. *Computers & Education*, 87, 131–141.

R

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

- Resch, C., Keulers, E., Martens, R., Van Heugten, C., & Hurks, P. (2018). Does strategy instruction on the Rey-Osterrieth Complex Figure task lead to transferred performance improvement on the Modified Taylor Complex Figure task? A randomized controlled trial in school-aged children. *Clinical Neuropsychologist*. Advance online publication. doi:10.1080/13854046.2018.1448438
- Resing, W. C. M. (1993). Measuring inductive reasoning skills: The construction of a learning potential test. In J. H. M. Hamers, K. Sijtsma, & A. J. J. M. Ruijsenaars (Eds.), *Learning potential assessment: Theoretical, methodological and practical issues* (pp. 219–241). Amsterdam/Berwyn, PA: Swets & Zeitlinger Inc.
- Resing, W. C. M. (1998). Intelligence and learning potential: Theoretical and research issues. In J. Kingma & W. Tomic (Eds.), *Conceptual issues in research on intelligence. Vol.5 of Advances in cognition and educational practice* (pp. 227–259). Stamford, Connecticut: JAI Press.
- 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. (2013). Dynamic testing and individualized instruction: Helpful in cognitive education? *Journal of Cognitive Education and Psychology*, 12(1), 81–95.
- Resing, W. C. M., Bakker, M., Pronk, C. M. E., & Elliott, J. G. (2016). Dynamic testing and transfer: An examination of children's problem-solving strategies. *Learning and Individual Differences*, 49, 110–119.
- Resing, W. C. M., Bakker, M., Pronk, C. M. E., & Elliott, J. G. (2017). Progression paths in children's problem solving: The influence of dynamic testing, initial variability, and working memory. *Journal of Experimental Child Psychology*, 153, 83–109.
- 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.
- Resing, W. C. M., Touw, K. W. J., Veerbeek, J., & Elliott, J. G. (2017). Progress in the inductive strategy use of children from different ethnic backgrounds: A study employing dynamic testing. *Educational Psychology*, 37(2), 173–191.

- 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.
- Resing, W. C. M., Tunteler, E., & Elliott, J. G. (2015). The effect of dynamic testing with electronic prompts and scaffolds on children's inductive reasoning: A microgenetic study. *Journal of Cognitive Education and Psychology*, 14(2), 231–251.
- 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.
- Resnick, M., Martin, F., Berg, R., Borovoy, R., Colella, V., Kramer, K. & Silverman, B. (1998). Digital manipulatives: new toys to think with. *Proceedings of the SIGCHI conference on human factors in computing systems* (pp. 281-287) (CHI '98). New York: ACM Press/Addison-Wesley Publishing Co.
- Rey, A. (1941). L'examen psychologique dans les cas d'encephalopathie traumatique [The psychological examination of cases of traumatic encephalopathy]. *Archives de Psychologie*, 28, 286–340.
- 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.
- Richardson, K., & Norgate, S. H. (2015). Does IQ really predict job performance? *Applied Developmental Science*, 19(3), 153–169.
- 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.
- Roncato, S., Sartori, G., Masterson, J., & Rumiati, R. (1987). Constructional apraxia: An information processing analysis. *Cognitive Neuropsychology*, 4, 113–129.
- Roth-Van der Werf, G. J. M., Resing, W. C. M., & Slenders, A. P. A. (2002). Task similarity and transfer of an inductive reasoning training. *Contemporary Educational Psychology*, 27, 296–325.

S

- Scherer, R., Greiff, S., & Hautamäki, J. (2015). Exploring the relation between time on task and ability in complex problem solving. *Intelligence*, 48, 37–50.
- 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>
- Senese, V. P., De Lucia, N., & Conson, M. (2015). Cognitive predictors of copying and drawing from memory of the Rey-Osterrieth complex figure in 7- to 10-year-old children. *The Clinical Neuropsychologist*, 29(1), 118–132.
- Serious Toys. (2011). Serious Toys. Retrieved from <http://www.serious-toys.com/>
www.serious-toys.com/index.html
- 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.
- Siegler, R. S. (1996). *Emerging minds: The process of change in children's thinking*. New York: Oxford University Press.
- Siegler, R. S. (2004). Learning about learning. *Merrill-Palmer Quarterly*, 50(3), 353–368.
- Siegler, R. S. (2007). Cognitive variability. *Developmental Science*, 10(1), 104–109.
- Siegler, R. S., & Svetina, M. (2002). A microgenetic/cross-sectional study of matrix completion : Comparing short-term and long-term change. *Child Development*, 73(3), 793–809.
- Siegler, R. S., & Svetina, M. (2006). What leads children to adopt new strategies? A microgenetic/cross sectional study of class inclusion. *Child Development*, 77(4), 997–1015.
- Siemens, G. (2013). Learning analytics: The emergence of a discipline. *American Behavioral Scientist*, 57(10), 1380–1400.
- Sijtsma, K. (1993). Psychometric issues in learning potential assessment. In J. H. M. Hamers, K. Sijtsma, & A. J. J. M. Ruijsenaars (Eds.), *Learning potential assessment: Theoretical, methodological and practical issues* (pp. 175–194). Amsterdam/Berwyn, PA, PA: Swets & Zeitlinger Inc.

- Simon, H. A. (1974). How big is a chunk? *Science*, 183, 482–488.
- Simon, H. A., & Kotovsky, K. (1963). Human acquisition of concepts for sequential patterns. *Psychological Review*, 70(6), 534–546.
- Stad, F. E., Van Heijningen, C. J. M., Wiedl, K. H., & Resing, W. C. M. (2018). Predicting school achievement: Differential effects of dynamic testing measures and cognitive flexibility for math performance. *Learning and Individual Differences*, 67, 117–125.
- Stad, F. E., Vogelaar, B., Veerbeek, J., & Resing, W. C. M. (2017). Young children's transfer of series completion in a dynamic test setting: Does cognitive flexibility play a role? *International Journal of School and Cognitive Psychology*, 4(4).
- Sternberg, R. J. (1985). *Beyond IQ: A triarchic theory of human intelligence*. New York: Cambridge University Press.
- Sternberg, R. J. (1997). The concept of intelligence and its role in lifelong learning and success. *American Psychologist*, 52(10), 1030–1037.
- Sternberg, R. J., & Gardner, M. K. (1983). Unities in inductive reasoning. *Journal of Experimental Psychology: General*, 112(1), 80–116.
- Sternberg, R. J., & Grigorenko, E. L. (2002). *Dynamic testing: The nature and measurement of learning potential*. New York: Cambridge University Press.
- Stevenson, C. E., Bergwerff, C. E., Heiser, W. J., & Resing, W. C. M. (2014). Working memory and dynamic measures of analogical reasoning as predictors of children's math and reading achievement. *Infant and Child Development*, 23, 51–66.
- Stevenson, C. E., Heiser, W. J., & Resing, W. C. M. (2013). Working memory as a moderator of training and transfer of analogical reasoning in children. *Contemporary Educational Psychology*, 38(3), 159–169.
- 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.
- Stevenson, C. E., Heiser, W. J., & Resing, W. C. M. (2016). Dynamic testing of analogical reasoning in 5-6 year olds: Multiple choice versus constructed-response training. *Journal of Psychoeducational Assessment*, 34(6), 550–565.
- 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.

Stringer, P. (2018). Dynamic assessment in educational settings: is potential ever realised? *Educational Review*, 70(1), 18–30.

T

- Taylor, E. M. (1959). *Psychological appraisal of children with cerebral defects*. Cambridge, MA: Harvard University Press.
- 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.
- 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.
- Tupler, L. A., Welsh, K. A., Asare-Aboagye, Y., & Dawson, D. V. (1995). Reliability of the rey-osterrieth complex figure in use with memory-impaired patients. *Journal of Clinical and Experimental Neuropsychology*, 17(4), 566–579.

U

- Ullmer, B., & Ishii, H. (2000). Emerging frameworks for tangible user interfaces. *IBM Systems Journal*, 39, 915–931.

V

- Van Gog, T., Kester, L., Nievelstein, F., Giesbers, B., & Paas, F. (2009). Uncovering cognitive processes: Different techniques that can contribute to cognitive load research and instruction. *Computers in Human Behavior*, 25(2), 325–331.
- Verhaegh, J., Fontijn, W. F. J., Aarts, E. H. L., Boer, L., & van de Wouw, D. (2011). A development support bubble for children. *Journal of Ambient Intelligence and Smart Environments*, 3(1), 27–35.
- 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.
- Verhaegh, J., Fontijn, W. F. J., & 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.
- Verhaegh, J., Fontijn, W. F. J., & Resing, W. C. M. (2013). On the correlation between children's performances on electronic board tasks and nonverbal intelligence test measures. *Computers & Education*, 69, 419–430.

- Verhaegh, J., Hoonhout, J., & Fontijn, W. F. J. (2007). Effective use of fun with a tangible interactive console. In *Proceedings of the 4th International Symposium on Pervasive Gaming Applications* (pp. 177–178).
- Verhaegh, J., Resing, W. C. M., Jacobs, A. P. A., & Fontijn, W. F. J. (2009). Playing with blocks or with the computer? Solving complex visual-spatial reasoning tasks: Comparing children's performance on tangible and virtual puzzles. *Educational & Child Psychology*, 19, 18–39.
- 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.

W

- Weisberg, R. W. (2015). Toward an integrated theory of insight in problem solving. *Thinking & Reasoning*, 21(1), 5–39.
- 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.

Y

- Yang, Y., Buckendahl, C. W., Juszkievicz, P. J., & Bhola, D. S. (2002). A review of strategies for validating computer-automated scoring. *Applied Measurement in Education*, 15(4), 391–412.

Z

- Zoanetti, N., & Griffin, P. (2017). Log-file data as indicators for problem-solving processes. In B. Csapó & J. Funke (Eds.), *The nature of problem solving: Using research to inspire 21st century learning* (pp. 177–191). Paris, France: OECD Publishing.

Appendix

APPENDIX I:

Further explanation of Grouping of Answer Pieces

To further explain the basic principle of grouping of answer pieces, consider the following example. A cleaner is working in a university building with multiple floors. All rooms have to be cleaned, so what would be an adaptive way of approaching this task? If the cleaner represents this as 200 separate rooms, with even more separate desks, windows, and trashcans to clean, they would surely be overwhelmed by the sheer magnitude of the job at hand. However, if it is represented as “one floor at a time”, the task already becomes more manageable. In addition, the risk of forgetting to clean a room drastically reduces by using such a system. It would not surprise anybody that alternatively cleaning one room at each floor would be far more error prone and less efficient. That is not to say that it could not lead to the same result, it would just take a lot more effort arriving at the final goal.

Similar to the cleaner in the university building, cognitive tasks consist of multiple steps and elements, and it can be helpful to group together the activities or pieces of information that are related to each other. We can, however, not directly look at a child’s representation of a problem, but can see the actions resulting from the representation. Through a rule-based algorithm it is possible to analyze and score certain patterns of these actions. Consider again the example of the university cleaner. By analyzing their actions, we can obtain information regarding his approach in the following way: we could install a sensor system in each door and detect when they go into a room. Our algorithm to see if they use the group “floors” would look something like this:

if (*the 1st room = on the first floor*) & if (*the 2nd room = on the first floor*) &
if (*the nth room = on the first floor*)
without cleaning a room on a different floor in between
the floor is grouped

So if all rooms on the same floor are cleaned in immediate succession of each other, they are considered grouped. However, if somewhere in between, a room on the second floor would be cleaned, the first floor rooms would no longer be considered grouped together. A similar strategy was used for detecting the grouping of answer pieces. The first step was determining which parts were adaptive to group together, based on a shared conceptual basis such as color, shape, etc. The second step consisted of writing the actual algorithm, which in this case was done using Microsoft Excel, because of its ease of use and general availability for all who had to work with the algorithm. For this thesis, this was applied for two different tasks; a series completion task and a complex figure task.

The series completion task used in our research was a pictorial task that used puppet figures. These figures consisted of eight pieces, which provided us with eight pieces which could be grouped together. The head is a separate piece which goes through separate transformations, and as such, cannot be grouped with the other pieces, leaving seven pieces open for useful grouping. In the example, three groups of pieces emerge; the arms, the legs, and the body. This would produce an algorithm as follows:

if (n^{th} piece = *an arm*) & if ($(n+1)^{\text{th}}$ piece = *an arm*)
the arms are grouped

In practice, the algorithm had to be further defined, which sheds light on the different possibilities that were considered equally “grouped” in detecting the sequences of the placement of pieces:

if (n^{th} piece = *left arm*) & if ($(n+1)^{\text{th}}$ piece = *right arm*)
OR if (n^{th} piece = *right arm*) & if ($(n+1)^{\text{th}}$ piece = *left arm*)
the arms are grouped

Similarly, the legs could be checked:

if (n^{th} piece = *left leg*) & if ($(n+1)^{\text{th}}$ piece = *right leg*)
OR if (n^{th} piece = *right leg*) & if ($(n+1)^{\text{th}}$ piece = *left leg*)
the legs are grouped

The body could be checked in a similar fashion, although the formula would be a bit more extensive because it consists of more pieces:

if (n^{th} piece = left body part) & if ($(n+1)^{\text{th}}$ piece = middle body part) &
 if ($(n+2)^{\text{th}}$ piece = right body part)
 OR if (n^{th} piece = left body part) & if ($(n+1)^{\text{th}}$ piece = right body part) &
 if ($(n+2)^{\text{th}}$ piece = middle body part)
 OR if (n^{th} piece = right body part) & if ($(n+1)^{\text{th}}$ piece = middle body part) &
 if ($(n+2)^{\text{th}}$ piece = right body part)
 OR if (n^{th} piece = right body part) & if ($(n+1)^{\text{th}}$ piece = left body part) &
 if ($(n+2)^{\text{th}}$ piece = middle body part)
 OR if (n^{th} piece = middle body part) & if ($(n+1)^{\text{th}}$ piece = left body part) &
 if ($(n+2)^{\text{th}}$ piece = right body part)
 OR if (n^{th} piece = middle body part) & if ($(n+1)^{\text{th}}$ piece = right body part) &
 if ($(n+2)^{\text{th}}$ piece = left body part)
 the body is grouped

In the example it can be seen that the actual order of placement within the group is considered irrelevant in the scoring of whether pieces are grouped or not. In the same way, it is considered irrelevant in which order the child places the groups; it did not matter to the scoring of the grouping of answer pieces whether the children would start with the legs, arms, body, or head. The actual number of groups into which the answer could be divided differed per item and was based on the transformations, colors, patterns, and “puppet anatomy”. A detailed account of the groups of pieces that were discerned can be found in *Chapter 4*. The possible number of groups for items ranged between two and five. For each item, the number of groups children placed was divided by the maximum number of groups that could be placed in the item.

Similarly, a scoring method for the grouping of answer pieces was developed for the complex figure drawing. Here, grouping of answer pieces was based on the sequence of drawing the elements of the complex figure, and the groups consisted of the lines that shared a configurational or relational basis. For example, lines that formed a large rectangle together would be easier to draw if they were in fact remembered as a large rectangle, as opposed to a set of separate lines. So, if the lines of the rectangle were drawn in immediate succession of each other, without drawing any other lines in between, the rectangle was considered grouped. To enable the scoring of the

grouping of lines, first the smallest units (lines) within the complex figure were discerned, leading to a total of 56 lines for the figure. Next, 24 groups of lines were discerned within the figure. As an example, part of the complex figure is provided in Figure 1.

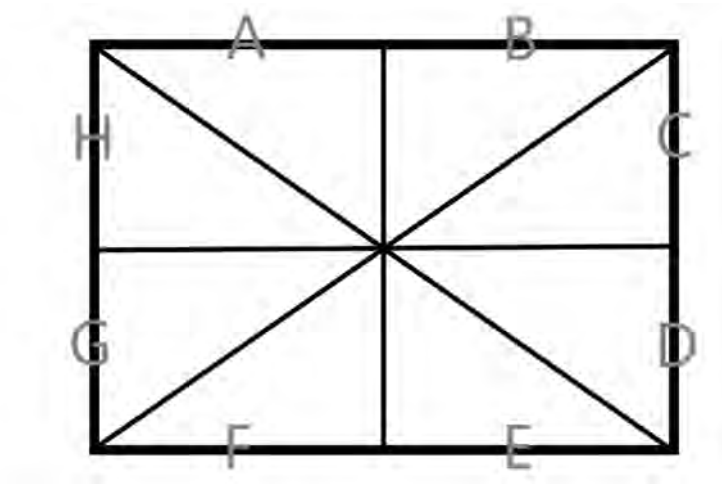


Figure 1. Example of the lines that were discerned for the grouping of answer pieces within a figure drawing.

Considering the part of the complex figure drawing as displayed in Figure 1, it can be seen that the smallest lines that were discerned were also connected to the other elements in the figure. In this case, the rectangle consist of 8 lines. So if these eight lines (A-H) were drawn in immediate succession of each other within the sequence of drawing, irrespective of the order in which they were drawn, they were considered grouped together. However, if a child would draw A, B, C, and would then continue by drawing the horizontal axis in the figure, the rectangle would no longer be considered grouped.

As can be expected, the groups that were discerned in this extensive geometric design, were in large part, simple geometric figures part of the design of the complex figure. *Chapter 5* contains a more detailed account of the groups that were discerned within the complex figure drawing. Similar to the scoring in the puppet figure, the score was divided by the total number of lines drawn, to obtain the proportion of grouped lines, and correct for the overall amount of lines drawn (lines that are not drawn, cannot be grouped).

Summary in Dutch

Samenvatting in het Nederlands

Een belangrijk deel van de cognitieve ontwikkeling van kinderen vindt plaats op school. Wanneer er vragen ontstaan over het cognitief functioneren van een kind, of de schoolse vooruitgang niet verloopt zoals die van leeftijdsgenoten, worden veelal statische tests ingezet om de cognitieve vaardigheden van het kind te onderzoeken (Elliott, Resing, & Beckmann, 2018; Resing, 2016). Onder statische tests worden traditionele cognitieve onderzoeksmiddelen verstaan die op één moment in de tijd worden afgenomen en waarbij geen feedback of training wordt geboden. Van deze statische tests wordt echter beweerd dat zij beperkte informatie bieden over de mogelijke redenen voor succes of falen op de test (Elliott, 2000; Elliott, Grigorenko, & Resing, 2010). Ze bieden weinig informatie over de individuele verschillen in processen die kinderen inzetten om cognitieve taken op te lossen (Richard & Zamani, 2003). Dynamisch testen is ontwikkeld om de nadelen van statische tests te kunnen ondervangen. Deze vorm van testen beoogt om het leerpotentieel van een kind in kaart te brengen. Hiervoor ontvangt het kind training of feedback tijdens de test. Sommige vormen van dynamisch testen zijn gericht op het in kaart brengen van de processen bij het oplossen van een taak, met als doel om informatie te verkrijgen over de oplossingsstrategieën die kinderen inzetten tijdens het werken aan een cognitieve taak (Resing, 2013). Het voornaamste doel in deze dissertatie was om te onderzoeken welke informatie kan worden verkregen met behulp van procesgeoriënteerd dynamisch testen met een theorie-gestuurde methode van procesanalyse. Daarbij was in het bijzonder aandacht voor de specifieke eigenschappen van een nieuwe geautomatiseerde methode om de oplossingsprocessen van kinderen tijdens het oplossen van cognitieve taken in kaart te brengen.

In *Hoofdstuk 1* is de theoretische achtergrond beschreven die relevant is voor de in deze dissertatie beschreven studies. In deze dissertatie is gebruik gemaakt van de ‘graduated prompts’ techniek, een specifieke vorm van dynamisch testen. Deze testen volgen een pretest/training/posttest opzet en kunnen worden ingezet om het minimum aantal prompts of hints te vinden dat een kind nodig heeft om de taak op te kunnen lossen (Campioni & Brown, 1978; Resing, 1993, 2000). Tijdens de training wordt hulp geboden in de vorm van hints of prompts, die in een van tevoren bepaalde hiërarchie worden aangeboden wanneer het een kind niet lukt om de taak zelfstandig correct te beantwoorden. Het aantal items dat een kind correct kan beantwoorden op de posttest en het aantal hints dat een kind nodig heeft gehad tijdens de training worden beiden gezien als redelijk accurate indicatoren van het leerpotentieel van een kind (Caffrey, Fuchs, & Fuchs, 2008; Resing & Elliott, 2011). Eerder onderzoek heeft laten zien dat dynamische testen aanvullende verklaarde variantie bieden in de voorspelling van schoolprestaties vergeleken met statisch testen (Elliott et al., 2018). Echter, procesgeoriënteerd dynamisch testen gecombineerd met computer geautomatiseerde scoring en graduated prompts training, zoals gehanteerd in deze dissertatie, is een relatief nieuwe ontwikkeling.

Voor het onderzoeken van oplossingsstrategieën zijn verschillende methoden beschreven in de literatuur (Ericsson, 2003; Tenison, Fincham, & Anderson, 2014), die elk een bepaalde compromis bieden tussen de accuratesse van de meting, de invloed die het heeft op het gedrag van de proefpersoon en het gebruiksgemak voor de tester (Tenison et al., 2014). Door technologische innovatie, zoals gebruiksvriendelijkere hardware en krachtigere softwaretoepassingen, zijn nieuwe mogelijkheden ontstaan voor het registreren van data en het analyseren van de acties die in deze data zijn gevat (Verhaegh, Fontijn, Aarts, & Resing, 2013). ‘Tangible user interfaces’ (TUI’s) bestaan uit tastbare, fysiek manipuleerbare materialen die zijn uitgerust met sensoren, zodat de acties die hierop worden uitgevoerd gemonitord kunnen worden. Zij bieden een meer natuurlijke werkvorm voor kinderen, waardoor kinderen op de taken natuurlijker oplossingsgedrag laten zien. Naast de interface, is er voor het meten van het oplossingsproces ook een systeem nodig om de data te vertalen in begrijpelijke informatie of scores, door een bepaalde interpretatie hieraan te geven. In deze dissertatie

is een theorie-gestuurde methode voor het onderzoeken van de oplossingsprocessen gehanteerd, in sommige studies in combinatie met een TUI.

De geautomatiseerde vorm van scoring die in deze dissertatie gebruikt is, '*Grouping of Answer Pieces*' (GAP), is gebaseerd op het detecteren van het groeperen van onderdelen van het antwoord op de taak. Hierbij werd gedetecteerd wanneer onderdelen die een conceptuele relatie hebben (bijvoorbeeld dezelfde transformatie of kleur) bij elkaar werden geplaatst in de volgorde van de antwoordstappen.

De studie beschreven in *Hoofdstuk 2* had als voornaamste doel de nieuwe maat voor de analyse van oplossingsprocessen (GAP) in seriële redeneertaken te evalueren bij kinderen van 8 jaar oud. De verwachting was dat de GAP-maat informatie zou bieden over de manier waarop kinderen de taak representeerden. Bij het oplossen van een taak wordt alle informatie over de taak in een interne representatie opgeslagen. Deze representatie is in de literatuur beschreven als een cruciaal onderdeel van het oplossingsproces voor de taak en bepaalt welke strategie wordt ingezet om de taak uiteindelijk te volbrengen. Om de informatie over de taakoplossingsprocessen van kinderen te kunnen verzamelen en natuurlijk gedrag van de kinderen uit te lokken, werd de taak afgenomen op een TUI. De GAP-maat bleek een voorspeller voor prestatie bij het oplossen van seriële redeneertaken. Daarnaast bleken verbalisaties over de gebruikte strategieën en tijdmaten voorspellende maten voor prestatie op de seriële redeneertaak. In dit artikel werd ook onderzocht welke factoren bijdragen aan succes op individuele items. Hierbij was te zien dat, afhankelijk van de moeilijkheidsgraad van het item, de GAP-maat de enige procesmaat was die bijdroeg aan de voorspelling van het succesvol oplossen van de taak. De GAP-maat was niet gerelateerd aan verbalisaties over de gebruikte strategieën of aan tijd maten, wat aangeeft dat het een uniek onderdeel van het oplossingsproces meet. Stabiele inzet van het groeperen van informatie was gerelateerd aan betere oplossingsvaardigheden in de seriële redeneertaak. Daarentegen was bij de verbalisaties over de gebruikte strategieën meer variabel gebruik juist gerelateerd aan betere prestaties op de seriële redeneertaak. Samen ondersteunen deze bevindingen het idee dat de verschillende maten voor oplossingsprocessen verschillende aspecten van het oplossingsproces meten.

In *Hoofdstuk 3* is onderzocht wat het effect was van het hanteren van een pretest op de resultaten op de posttest in een dynamische seriële redeneertaak. Hierbij kreeg de helft van de kinderen een pretest, een graduated prompts training en een posttest, de andere helft van de kinderen kreeg alleen een graduated prompts training en een posttest. De effecten van het afnemen van een pretest in seriële redeneertaken werd onderzocht op zowel het gebied van prestaties, als de oplossingsprocessen van kinderen op de posttest. In dit onderzoek participeerden kinderen van 7 en 8 jaar oud. Er werden geen verschillen gevonden tussen de kinderen waarbij een pretest was afgenomen en kinderen waarbij dat niet het geval was. De onderzochte gebieden waren de oplossingsvaardigheden op de posttest, de processen die werden ingezet tijdens het oplossen van de taken zoals gemeten met de GAP-maat, verbalisaties over de gebruikte strategieën, tijdmaten en het aantal hints dat kinderen nodig hadden tijdens de training. Aanvullende analyses lieten zien dat de procesmaten bijdroegen aan de voorspelling van taaksucces. Het wel of niet hanteren van een pretest leek geen verschil te maken voor de psychometrische eigenschappen van de posttest, wat aangaf dat de beslissing over of een pretest gehanteerd moet worden kan worden genomen vanuit de onderzoeksvraag die op dat moment belangrijk is in plaats van door psychometrische aspecten.

Het onderzoek dat is beschreven in *Hoofdstuk 4* had als doel om bij 7 en 8-jarige kinderen de processen die ingezet werden bij het oplossen van seriële redeneertaken te onderzoeken en hierbij de invloed van training op de processen en uitkomsten te analyseren. De helft van de kinderen ontving een graduated prompts training, de andere helft alleen herhaalde afname. De processen die kinderen inzetten om de taken op te lossen werden gemeten met de GAP-maat, verbalisaties over de gebruikte strategieën en de tijd die de kinderen nodig hadden om de items te beantwoorden. Getrainde kinderen verbeterden meer in hun prestaties op de seriële redeneertaken van pre- naar posttest dan ongetrainde kinderen. Op de procesmaten was het effect van training het duidelijkst te zien op de verbalisaties over de gebruikte strategieën, die verbeterden door de training. Op de GAP-maat was geen verschil te zien in het niveau van getrainde en ongetrainde kinderen, maar er waren wel verschillen in de verdeling van de verschillende GAP categorieën tussen getrainde kinderen en kinderen die niet getraind waren. Op de tijd die kinderen nodig hadden om de taken te volbrengen waren

geen differentiële effecten te zien als resultaat van de training of herhaalde afname. De procesmaten waren gerelateerd aan prestatie op de dynamische seriële redeneertaak en aan rekenen en begrijpend lezen zoals gemeten met het Cito leerlingvolgsysteem. De hoeveelheid hulp die kinderen nodig hadden tijdens de training was de beste voorspeller voor schoolprestaties.

In *Hoofdstuk 5* is procesgeoriënteerd dynamisch testen binnen een ander domein toegepast. Hiervoor werd opnieuw een pretest/training/posttest ontwerp gehanteerd, waarbij de helft van de kinderen training kreeg en de andere helft alleen herhaalde afname. Een visuele complexe figuur taak werd afgenomen om de effecten van training op de processen en prestaties van 7 en 8-jarige kinderen te evalueren. Voor deze studie is een graduated prompts training ontwikkeld, die voornamelijk tot doel had de kinderen te leren het complexe figuur beter te organiseren. Om het groeperen van de onderdelen van het antwoord te kunnen monitoren, werd de GAP-maat aangepast voor deze taak. De training leidde tot meer vooruitgang in de prestaties van kinderen en tot meer groepeergedrag tijdens het tekenen van het complexe figuur. Deze bevinding gaf aan dat de processen die kinderen inzetten om het complexe figuur te tekenen veranderden als resultaat van de graduated prompts training. Alleen getrainde kinderen kwamen tot het hoogste niveau van organisatie van het complexe figuur op de posttest.

In *Hoofdstuk 6* zijn de studies samengevat en de theoretische en praktische implicaties hiervan beschreven. Allereerst werd geconcludeerd dat procesgeoriënteerd dynamisch testen mogelijk meer informatie kan bieden over de leerbehoeften van een kind dan traditionele, statische tests. In deze dissertatie werd gevonden dat procesgeoriënteerd dynamisch testen waardevolle informatie kan bieden als aanvulling op de prestaties van een kind en informatie kan bieden over de effecten van training op de processen die kinderen inzetten om taken op te lossen.

De meeste studies in deze dissertatie hanteerden een dynamisch testontwerp, waarbij de graduated prompts methode werd gebruikt in de trainingsfase. Dynamisch testen leidde tot significante verbetering in de prestaties van kinderen op zowel de seriële redeneertaak als de complex figuur taak. Dit is in lijn met bevindingen uit het verleden waarbij de graduated prompts methode succesvol werd ingezet bij serieel redeneren (Resing, Touw, Veerbeek, & Elliott, 2017; Resing, Xenidou-Dervou, Steijn, & Elliott, 2012; Stad, Vogelaar, Veerbeek, & Resing, 2017), analogisch redeneren

(Stevenson, Heiser, & Resing, 2013) en taalvaardigheden (Camilleri & Botting, 2013; Hasson, Dodd, & Botting, 2012). Dynamische testmaten leken bovendien, in lijn met wat in eerder onderzoek is gevonden, tot meer verklaarde variantie te leiden dan het hanteren van statische testmaten. Het aantal hints dat kinderen nodig hadden tijdens de trainingsfase leek de beste voorspelling te bieden voor de schoolprestaties van kinderen. In hoofdstuk 3 is het effect van de inzet van een pretest specifiek onderzocht. Hieruit werd geconcludeerd dat de beslissing over de inzet van een pretest in eerste instantie kan worden gebaseerd op de onderzoeksvraag voor het betreffende onderzoek en niet af hoeft te hangen van overwegingen over de psychometrische eigenschappen van het hanteren van een pretest.

De verschillende procesmaten die benut werden tijdens het onderzoek leken allemaal een ander aspect van het taakoplossingsproces te meten. Verbalisaties leken vooral strategiegebruik weer te geven zoals is beschreven door Siegler (1996, 2007), die strategiegebruik omschreef als variabel, zowel binnen als tussen taken. Deze variabiliteit was gerelateerd aan betere prestaties, wat aangeeft dat succesvol oplossingsgedrag meer verbonden is aan de vaardigheid om flexibel tussen strategieën te kunnen wisselen. Onder invloed van training verbeterden kinderen in het geven van verbalisaties. Tijdsmaten waren in mindere mate gecorreleerd aan prestaties en leken niet beïnvloed te worden door training, wat een resultaat kan zijn van de moeilijkheidsgraad van de taken. Eerder onderzoek heeft aangegeven dat de relatie tussen prestatie en tijdsmaten wordt beïnvloed door de moeilijkheidsgraad van taken.

Het groeperen van oplossingsstappen van het antwoord (GAP) was een andere procesmaat, waarvan werd verwacht dat deze een indicator was voor de manier waarop kinderen de taak representeren, door het inzetten van strategieën om de informatie over de taak efficiënter op te slaan door het gebruik van “chunks” (Halford, Wilson, & Phillips, 1998). Deze maat was gerelateerd aan accuratesse op de seriële redeneertaak, maar was niet gerelateerd aan de andere procesmaten, wat aangeeft dat de maat mogelijk unieke informatie kan bieden over de taakoplossingsprocessen van kinderen. Bij de GAP-maat was stabiele inzet gerelateerd aan betere prestaties, wat overeenkomt met de theorie van Sternberg (1985), die representatie omschreef als een metacomponent voor het probleemoplossingsproces. De GAP-maat was in de studies beschreven in deze dissertatie consistent een gematigde

voorspeller voor prestatie, wat aangaf dat de representatie zoals gemeten met de GAP-maat gerelateerd is aan succesvol taak oplossingsgedrag (Hunt, 1980; Pretz, Naples, & Sternberg, 2003; Robertson, 2001).

Om te onderzoeken of de GAP-maat ook kon worden aangepast voor inzet in andere taken, werd deze maat aangepast aan de complexe figuur taak, gecombineerd met een graduated prompts training die erop gericht was om kinderen te ondersteunen in het vormen van een georganiseerde representatie van het figuur. Hierbij zorgde de training voor significante verschillen in zowel de prestaties op de posttest als op de GAP-maat ten opzichte van herhaalde afname.

In lijn met recent onderzoek, werd in de studies van deze thesis ook gevonden dat de eigenschappen van de taak in grote mate bijdragen aan de relatie tussen de procesmaten en prestatie. Wanneer op het niveau van individuele items werd gekeken, waren de taakeigenschappen de voornaamste verklarende factor voor succes op de taak. Dit geeft aan dat wanneer de processen die bijdragen aan het oplossen van een taak worden onderzocht, het van belang is om naar de factoren te kijken die bijdragen op het niveau van individuele items en hierbij ook rekening te houden met complexe en interactieve relaties.

Voor het onderzoeken van de cognitieve vaardigheden van kinderen bieden computers een scala aan mogelijkheden, op het vlak van het vastleggen en scoren van data die is verkregen tijdens het testen in een gecomputeriseerde of pen-en-papier type taak. Dit kan de werkbelasting van het onderzoek voor de afnemer verlagen, terwijl ook informatie over de oplossingsprocessen die kinderen inzetten relatief gemakkelijk kan worden verkregen. Door het gebruik van TUI's kunnen de voordelen van computers worden gecombineerd met die van fysieke, tastbare materialen voor de ontwikkeling van kinderen en het ontlocken van natuurlijk gedrag bij kinderen. Deze gecomputeriseerde systemen maken het mogelijk dat kinderen zelfstandig aan taken werken en geautomatiseerde feedback en training kunnen ontvangen. Het geeft ook de mogelijkheid om door middel van theorie-gestuurde scoring informatie te verkrijgen over de processen die kinderen inzetten tijdens het oplossen van de taken.

Er dient wel een kanttekening te worden gemaakt bij het hanteren van procesgeoriënteerd testen. Hier is nog relatief weinig over bekend en het is niet altijd gemakkelijk toe te passen op nieuwe taken, mede doordat

de relaties tussen de data soms interactief zijn en moeilijk te interpreteren. Ook is er meer onderzoek nodig voor het verbinden van specifieke betekenis aan procesmaten en deze te verbinden met aanbevelingen voor interventie. Toch biedt gecomputeriseerd procesgeoriënteerd dynamisch testen een aantal unieke mogelijkheden voor assessment in het onderwijs. Zo kan het leerpotentieel van kinderen en de oplossingsprocessen die kinderen inzetten worden onderzocht, terwijl zij zelfstandig met de technologie werken. Op deze manier kan de werklast voor leerkrachten lager gehouden worden. Het benutten van de mogelijkheden die de technologie die de 21ste eeuw ons biedt mag dan een uitdaging zijn, het biedt ook nieuwe mogelijkheden voor het onderzoeken en ontwikkelen van het individuele leerpotentieel van kinderen.

Curriculum Vitae

Jochanan Veerbeek was born on 13 March 1985 in Zwijndrecht. He obtained his secondary school *havo* diploma in 2003 from Develsteincollege in Zwijndrecht. After obtaining his diploma, he worked as an audio engineer and as a technical equipment support engineer. He decided to continue his education in 2008 at the Hogeschool Leiden, where he successfully finished the propaedeutic stage in Music Therapy in 2009. From here, he continued at Leiden University where he obtained a Bachelor's degree in Psychology in 2012 and a Master's degree in Child and Adolescent Psychology in 2014. He wrote his Master's thesis under the supervision of Professor Wilma Resing, who introduced him to the concept of dynamic testing. After obtaining his Master's degree, he continued on this subject as an external PhD candidate, with the encouragement and support of Wilma Resing as his promotor. Meanwhile, he combined working at Leiden University as a thesis supervisor, with working in practice with children with learning disabilities at the Feuerstein Centrum Nederland in Amsterdam. Currently, he is still working at the Feuerstein Centrum Nederland as a mediator and assessor using dynamic assessment.

