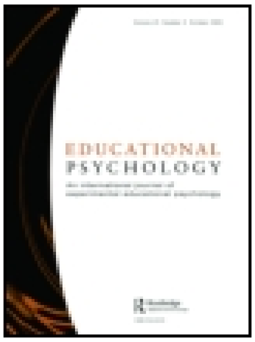

Routledge
Taylor & Francis Group

Educational Psychology

An International Journal of Experimental Educational Psychology

ISSN: 0144-3410 (Print) 1469-5820 (Online) Journal homepage: https://www.tandfonline.com/loi/cedp20

Children's solving of 'Tower of Hanoi' tasks: dynamic testing with the help of a robot

Wilma C. M. Resing, Bart Vogelaar & Julian G. Elliott

To cite this article: Wilma C. M. Resing, Bart Vogelaar & Julian G. Elliott (2019): Children's solving of 'Tower of Hanoi' tasks: dynamic testing with the help of a robot, Educational Psychology, DOI: 10.1080/01443410.2019.1684450

To link to this article: https://doi.org/10.1080/01443410.2019.1684450

Published online: 20 Nov 2019.

EDUCATIONAL PSYCHOLOGY
https://doi.org/10.1080/01443410.2019.1684450

Children's solving of 'Tower of Hanoi' tasks: dynamic testing with the help of a robot

Wilma C. M. Resing[a], Bart Vogelaar[a] and Julian G. Elliott[b]

[a]Department of Developmental and Educational Psychology, Leiden University, Leiden, The Netherlands; [b]School of Education, Durham University, Durham, United Kingdom

ABSTRACT

The present study investigated the usefulness of a pre-programmed, teleoperated, socially assistive peer robot in dynamic testing of complex problem solving utilising the Tower of Hanoi. The robot, in a 'Wizard of Oz' setting, provided instructions and prompts during dynamic testing to children when they had to solve 3 D Tower of Hanoi puzzles. Participants were 37 second grade 8-year-old children, of whom half received graduated prompts training between pre-test and post-test, delivered by the robot, and half did not. It was found that children's progression in task accuracy varied considerably, depending on whether or not children were trained in solving Tower puzzles. Trained children showed greater progression in the number of Tower problems that they could solve accurately, made considerably fewer steps, although the Tower puzzles increased quickly in difficulty level. The mean completion time of trained children decreased at a slower rate than that of the untrained children, but both groups of children took considerably more time to think and plan ahead before they started the solving process. Only moderate relations with planning behaviour were found. In general, the study revealed that computerised dynamic testing with a robot as assistant has much potential in unveiling children's potential for learning and their ways of tackling complex problems. The advantages and challenges of using a robot in educational assessment were discussed.

ARTICLE HISTORY
Received 4 December 2018
Accepted 20 October 2019

KEYWORDS
Dynamic testing; complex reasoning; Tower of Hanoi; computerised dynamic testing; planning

Introduction

New technology in education and assessment

New educational technology involving the use of computers, tablets and even robots in instruction or educational assessment procedures has been developed rapidly, and is becoming increasingly widespread (Baxter, Ashurst, Read, Kennedy, & Belpaeme, 2017; Benitti, 2012; Hong, Huang, Hsu, & Shen, 2016). Chin, Hong, and Chen (2014), for

CONTACT Wilma C. M. Resing resing@fsw.leidenuniv.nl Department of Psychology, Faculty of Social Sciences, Section Developmental and Educational Psychology, Leiden University, P.O. Box 9555, 2300 RB Leiden, The Netherlands

example, have studied the usability of robots as instructional tools for transmitting knowledge in education. Using robots in the classroom or as a form of support in an assessment context seems promising, as prior research has revealed such tools can have a positive influence on children's motivation when tackling cognitive tasks (André et al., 2014; Deublein, 2018; Tanaka, Cicourel, & Movellan, 2007). Young children have shown their ability to learn from a peer or supportive tutoring robot in several cognitive domains such as vocabulary, (second) language learning, mathematics, science, and in the more general domain of thinking skills (Belpaeme, Kennedy, Ramachandran, Scassellati, & Tanaka, 2018; Chang, Lee, Chao, Wang, & Chen, 2010; Hussain, Lindh, & Shukur, 2006; Jones & Castellano, 2018; Kanero et al., 2018; Moriguchi, Kanda, Ishiguro, Shamida, & Itakura, 2011; Movellan, Eckhardt, Virnes, & Rodriguez, 2009; Sullivan, 2008; Tanaka & Matsuzoe, 2012). Outcomes from these studies have revealed that children showed positive interactions when learning together with a robot. In their study, André et al. (2014) also reported robots influencing children's behaviour positively when mental arithmetic tasks had to be solved.

Robots used in education have adopted different formats and are functioning more or less independently at an increasingly sophisticated level. According to Mubin, Stevens, Shahid, Al Mahmud, and Dong (2013), they can be classified on the basis of the extent to which they can operate autonomously. It is possible to distinguish between the following types of robot: as a tool (a technology aid, for example in solving 3-D tasks), a peer robot (often presented as a knowledgeable peer who guides the child in a learning process), or a tutor (providing direct support with the curriculum, sometimes in the form of a robot teaching assistant). The current study examined the use of a teleoperated socially assistive peer robot that provided instructions during multiple assessment sessions.

Dynamic testing

Educational assessment of children's cognitive strengths and weaknesses on a one-to-one basis is often quite time-consuming. To support educational or school psychologists, and teachers, robotic assessment applications may be particularly useful, specifically in relation to dynamic testing. Educational and school psychologists often make use of conventional, static tests, considered as efficient tools for measuring already developed knowledge and abilities at a single point in time. Dynamic test measures, on the other hand, are designed to assess abilities that are yet-to-be developed (Elliott, Grigorenko, & Resing, 2010; Elliott, Resing, & Beckmann, 2018; Sternberg & Grigorenko, 2002). An important aspect of a dynamic testing procedure is the provision of feedback and tailored training between testing sessions (e.g. Elliott et al., 2010). In so doing, dynamic testing aims to focus on children's potential for learning, rather than their past learning outcomes. However, this approach has its challenges. As dynamic tests typically include several assessment points, their administration can be time-consuming. Moreover, studying children's potential for learning in detail can easily result in a lack of standardisation due to the interactive nature of the assessment procedure. As providing instruction and feedback, sometimes in the form of prompts and scaffolds, is key to the process of dynamic testing, the use of a supporting peer robot to administer these tests would seem to have much potential. Unlike conventional computerised forms of assessment, dynamic testing with a robot does not require the child to manipulate a mouse while watching a computer screen. Instead, they can freely manipulate the three-dimensional tangibles of the task to be administrated (e.g. Verhaegh, Fontijn, Aarts, & Resing, 2013). Such support can prove helpful in motivating children to do their best to solve the tasks presented to them.

The dynamic testing approach utilised in our study using a pre-test-training-post-test format, provides guided cognitive and metacognitive instruction and observation of the task solving process of each individual child. Central in this graduated prompts approach is the incorporation of structured feedback and, where possible, tailored assistance into the training procedures (Elliott et al., 2010, 2018; Grigorenko, 2009; Jeltova et al., 2011). In the present study, children had to solve tangible, three-dimensional Tower-of-Hanoi tasks. Our approach permitted recording of the duration, the nature of each solving step, as well as any solving activities that were not allowed within the framework of the Tower-of-Hanoi task of the children in a quasi-natural, seamless setting (Föβl, Ebner, Schön, & Holzinger, 2016). The training phase existed of an hierarchical step-by-step provision of cognitive and metacognitive prompts based on former task analyses (e.g. Kotovsky, Hayes, & Simon, 1985; Welsh, 1991; Welsh & Huizinga, 2001), and were provided in accordance with the child's perceived needs. The prompts provided as little help as possible to enable accurate task completion. This was geared to support children in solving the training tasks as independently as possible and to heighten motivation.

Progression in independent task performance after the training and the total number of prompts needed during training were both considered to be indices of a child's potential for learning. In the current study, we programmed this hierarchical step-by-step prompts and scaffolding procedure to ensure that the training procedure was both adaptive and structured.

Dynamic testing by computer and robot

Several researchers have focussed on the value and effects of computerised dynamic testing in educational settings. Tzuriel and Shamir (2002) reported that children, when dynamically assisted by a computer, showed more progression in seriational reasoning than children who were dynamically assessed by a human examiner. In contrast, studies by Stevenson, Touw, and Resing (2011) and Resing, Steijn, Xenidou-Dervou, Stevenson, and Elliott (2011), in which children solved analogies, did not reveal accuracy differences in computerised dynamic testing versus dynamic testing by a human assessor. According to Resing et al. (2011), Resing, Touw, Veerbeek, & Elliott (2017) and Resing, Xenidou-Dervou, Steijn, and Elliott (2012), computerised dynamic testing provided more fine-grained information regarding the variability in solving processes of individual children. Some authors in social robots research concluded in their research that robot-based instruction methods could result in similar, and more natural, effects as instruction provided by humans (e.g. Baxter, et al., 2017; Brown & Howard, 2013). Serholt, Basedow, Barendregt, and Obaid (2014) further noted that, although the children in their study asked the human instructor more often for help, the children were very good in utilising instructions from a support robot. Moreover, a recent study into dynamic testing of inductive reasoning revealed the potential usefulness of being tested by a robot (Resing, Bakker, Elliott, & Vogelaar, 2019). In this study, it was found that children who were dynamically tested by a robot improved more in inductive reasoning than their peers who only completed the pre-test and post-test.

On the basis of these findings, it was expected that computerised assisted instruction with the support of a robot would offer promising new possibilities for dynamic testing of complex reasoning. The use of a small table robot as an instructor could potentially provide more adaptive prompts and scaffolding procedures, and create a natural, authentic assessment environment (Huang, Wu, Chu, & Hwang, 2008; Khandelwal, 2006, Khandelwal & Mazalek, 2007), in particular when children have to solve complex reasoning or problem-solving tasks. Although the research findings above seem to bode well for the use of computerised (dynamic) testing, such an assumption requires closer inspection. In particular, when the tasks involved are three-dimensional but presented as two-dimensional on a screen, unforeseen problems may arise. A trial-and-error approach has been associated with poorer performance on a two-dimensional computer representation of various tasks, when compared with three-dimensional task versions presented by a human examiner (e.g. Noyes & Garland, 2003; Salnaitis, Baker, Holland, & Welsh, 2011).

In the current study, we utilised a dynamic and adapted three-dimensional version of the Tower of Hanoi (e.g. Welsh, 1991). For this task, in particular, representation on a computer screen was found to require more solution steps to achieve a successful outcome than when the child could manipulate the actual pieces of the Tower one by one in a traditional setting on the table (Noyes & Garland, 2003). Therefore, in comparison with presentation on a computer screen, robot assistance was considered to have the advantage that it allows the child to actively manipulate the tangible Tower objects in a natural setting.

Tower of Hanoi, executive functioning, and planning

A number of studies on dynamic testing make use of inductive reasoning tasks, such as (verbal or geometric) analogies, matrices, seriation or inclusion (e.g. Goswami, 2013; Resing et al., 2017; Resing & Elliott, 2011; Tzuriel, Isman, Klung, & Haywood, 2017; Vogelaar & Resing, 2018). In the current study, however, a newly adapted, dynamic version of the 'Tower of Hanoi' task was used. The origin and use of this task, in general, have been thoroughly described in Hinz, Klavžar, Milutionvić, and Petr (2018). The Tower of Hanoi task has often been used to measure systematic strategy use in children, because of its novel and abstract nature (Welsh, 1991). Within the cognitive domain, the Tower of Hanoi task is associated with measuring visual-spatial and complex problem-solving abilities in both adults (Kotovsky et al., 1985) and children (Kaufman, 2007; Klahr & Robinson, 1981). Besides these cognitive requirements, successful task solving has been seen as related to executive functioning, and planning in particular (Culbertson & Zillmer, 1998). Without planning it is hardly possible to accurately solve the more difficult items (Bishop, Aamodt-Leeper, Creswell, McGurk, & Skuse, 2001).

Planning is considered to be an important prerequisite for solving complex reasoning tasks and requires the inhibition of immediate actions. Some children immediately start to produce a problem solution (trial and error), instead of first trying to think over the best steps and strategy towards solving the task and can be unduly optimistic about their chance of succeeding without planning (Zimmermann, 2002). Moreover, children often make mistakes in complex and novel task situations if they do not use planning strategies accurately (Berg, Strough, Calderone, Meegan, & Sansone, 1997). Planning is considered to be a higher-order executive function and concerns the ability to think ahead, set goals, and anticipate problem-solving steps and strategy use (e.g. Lehto, Juujärvi, Kooistra, & Pulkkinen, 2003; Meltzer, 2018). In solving complex cognitive tasks, such as Tower of Hanoi tasks, it is often useful to tackle the task analytically, step-by-step, thinking ahead of future solving-steps (Welsh & Huizinga, 2001). Various researchers have used Tower of Hanoi tasks to study children's planning abilities when solving complex cognitive tasks (Carlson, Moses, & Claxton, 2004; Klahr & Robinson, 1981; Simon, 1975; Welsh, 1991; Welsh & Huizinga, 2005). These studies have led researchers to conclude that preliminary planning ability develops during the pre-school years (Carlson et al., 2004). In addition to measuring planning, the Tower of Hanoi is said to tap into other aspects of executive functioning, such as working memory and inhibition (Welsh, Satterlee-Cartmell, & Stine, 1999).

General aim of study and research questions

In the current study, a small table robot was utilised for a series of test and training sessions in which children repeatedly had to solve three-dimensional Tower of Hanoi tasks. Recent research suggests positive effects regarding children's engagement with robots in the school (e.g., André et al., 2014; Baxter et al., 2017; Belpaeme et al., 2018; Benitti, 2012; Beran & Ramirez-Serrano, 2010; Cha, Greczek, Song, & Matarić, 2017; Jones & Castellano, 2018; Kozima & Nakagawa, 2007; Libin & Libin, 2010; Moriguchi, Kanda, Ishiguro, Shimada, & Itakura, 2011; Movellan et al., 2009; Tanaka et al., 2007; Toh, Causo, Tzuo, Chen, & Yeo, 2016). Based on these findings, we sought to examine the potential of utilising a robot in a dynamic testing setting in young primary school children, both as a means to assess and record their outcome performance and changes in performance after the training, and to consider the impact of the robot's interaction with the children. The robot was controlled by an examiner who followed a pre-programmed computerised script, thereby functioning as a 'Wizard of Oz' figure (Dahlbäck, Jönsson, & Ahrenberg, 1993). It was expected that the robot, operating in a standardised but playful way, would enable us to examine the children's problem solving processes and learning progression (e.g. Resing & Elliott, 2011).

We sought to examine whether computerised dynamic testing, including a (teleoperated) robot as instructor, would result in systematic and controlled dynamic testing procedures. More specifically, we examined the effects of receiving instruction and training by a table robot on children's performance across assessment sessions, the role of executive functioning, and more specifically the planning behaviour of the children, before and after training.

Firstly, we considered the effects of the robot's training on children's responses to the Tower of Hanoi problems. It was expected that trained children would demonstrate both greater pre-test to post-test progression, and fewer steps in solving the Tower puzzles, than untrained children. These expectations were based on research findings regarding the effects of electronic ways of training procedures in dynamic testing, but with different tasks (e.g. Passig, Tzuriel, & Eshel-Kedmi, 2016; Resing et al., 2012; Tzuriel & Shamir, 2002; Wu, Kuo, & Wang, 2017). In some of these studies, children's solving of 3 D tasks on an electronic console was measured utilising sensor technology (Resing et al., 2017; Resing & Elliott, 2011).

Secondly, the effects of training on children's solving and pre-solving time, the time spent on preparing the task solution before executing any steps, were examined. It was expected that the time children utilised to solve all the tower tasks would increase across test sessions, particularly for the trained children (e.g. Resing & Elliott, 2011). Although the results of earlier studies have not always proven supportive of this hypothesis (e.g., Resing, Tunteler, & Elliott, 2015; Veerbeek, Hessels, Vogelaar, & Resing, 2017), we expected that task complexity would be of influence here. To further support this hypothesis, it was expected that training would positively influence children's pre-solving behavior, and that, as a consequence, after training, children would start to take action later than their unguided peers (e.g. Kossowska, & Necka, 1994; Resing et al., 2012), particularly because the Tower of Hanoi tasks are considered to be complex problem-solving tasks for children (e.g. Kaufman, 2007; Klahr & Robinson, 1981).

Thirdly, we evaluated the number of prompts children needed and the warnings (provided when children did not follow the solving rules of the Tower of Hanoi tasks) they received during the training, in relation to their progress from pre-to post-test and planning behaviour. We explored whether trained children would need fewer warnings for irregular solving behaviour than untrained children and whether the number of warnings given to the trained children were related to task accuracy in the pre-test phase, but not in the post-test phase.

Finally, we examined differences in the children's executive functioning (particularly, planning). We explored whether children with poorer planning skills, as reported by their teachers, would profit more from training, and therefore would show different progression lines than children with more advanced planning skills (e.g. Vogelaar, Bakker, Hoogeveen, & Resing, 2017).

Method

Participants

The participants in this study were 37 children from regular primary schools, 19 boys and 18 girls with a mean age of 94.57 months (*SD* = 7.51 months). Schools were selected on the basis of their willingness to participate. All children were in second grade of elementary school, at the end of the school year, and came from two primary schools in the Western part of the Netherlands. They were all from middle to higher range of socio-economic backgrounds. All children did not have any experience in working with a robot, or robotic programming. The testing was conducted by three

Table 1. Outline of the TOH study design.

Condition	Raven	Pre-test TOH	Training TOH	Post-test TOH
Experimental (trained)	x	X	X	X
Control (non-trained)	x	X	Practicing with different tasks	X

postgraduate students with teaching experience, who were trained extensively in the study procedures. For all children, Dutch was the first language spoken at school and at home. Data of three of the 40 initially participating children were excluded from the analyses, because they participated only in the pre-test session or were just playing with the test materials during pre-test, without showing any attention to the task instruction. Written parental permission after informed consent was obtained for all children before they started to participate in the study. All procedures, including the informed consent and the recruitment of participants, were reviewed and approved by the Institutional Committee on Ethics in Psychology.

Design

The study employed a pre-test-training-post-test control-group design with randomised blocking on the basis of children's scores on a general inductive reasoning test, the Raven Progressive Matrices (Raven, Raven, & Court, 2003) administered before the dynamic testing commenced (See Table 1 for an overview of the design). On the basis of the blocking, pairs of children, per school, were randomly assigned to one of two Condition groups: a 'training' and a 'non-training' group. All the children were given a pre and post-test with Tower of Hanoi problems. Children in the training condition received a 45-minute training session between the pre- and post-test sessions, whereas non-trained children completed other cognitive tasks, such as mazes and dot-to-dot tasks, in the same time period. The robot was present at the children's table during all test sessions, with the exception of the Raven test and the paper-and-pencil control task sessions for the non-trained group.

Materials

Raven progressive matrices

The Raven (Raven et al., 2003) was administered to provide an indication of the general level of the children's inductive reasoning ability. The test consists of 60, 3×3 figural matrix items in which one part is missing. From a number of alternatives, children had to generate the accurate answer by induction out of 6 alternatives. A split-half-coefficient was reported as a measure of the reliability of the test ($r = 0.91$; Raven et al., 2003).

Behaviour Rating Inventory of executive function

An adapted, Dutch version of the Behaviour Rating Inventory of Executive Function (BRIEF) (Gioia, Isquith, Guy, & Kenworthy, 2000; Huizinga & Smidts, 2010; Smidts & Huizinga, 2009) was completed by the teachers for all participating children. The questionnaire consists of 75 statements with three response options each and has been designed for describing the metacognitive and executive behaviours of children in the

Figure 1. Picture of the green robot (WittyWorxk, 2012).

age range of 5–17 years. This instrument includes eight subscales. In this study, only the subscale Planning and Organising and the General Executive Functioning Index were utilised, both measuring aspects of the executive functioning of a child as observed by their teacher (e.g. Smidts & Huizinga, 2009). The internal consistencies (Cronbach's α) of these two scales were reported as .91 and .97, respectively, (Huizinga & Smidts, 2010).

The robot

In this study, a 20 cm tall robot called 'Myro', depicted in Figure 1, was placed on a child's desk during the test sessions. This device has been developed by WittyWorxk (2012) and has an appearance similar to that of a friendly green owl. The robot was pre-programmed to speak, dance, move, show feedback with its eyes, and react to touch (head moving, turning around, dancing). Non-verbal behaviour included happy, sad, surprised, or neutral expression, as shown by the eyes (two colour displays and with different sizes). Nodding, head-shaking, and dancing (body movement or turning around) were possible in all directions. With its sensors, sounds and expression abilities, it was anticipated that the robot could successfully hold the children's attention and interact with them in a playful way.

The robot's stand-alone stage of development required the deployment of a Wizard of Oz setting. The examiner, sitting in the room out of the child's direct sight, served as the 'eyes and ears' of the robot. The robot was constructed with a camera, microphones, and touch sensors inside, so that the solving processes of the children could be filmed (only the voice and the movement of the hands of each child were filmed). This setting enabled the examiner to follow and analyse detailed aspects of the problem-solving behaviour of the child during the Tower of Hanoi test sessions. The Tower tasks of all parts of the dynamic test were simulated on the computer and the examiner had to 'mimic' the task exactly and at the very same moment as the child, following a strict if-then scenario, and thus causing the robot to respond accordingly. Thus, the robot was able to interact and give feedback at exactly the right time.

Tower of Hanoi: general

In this study, an adapted coloured tangible version of the 'Tower of Hanoi' task was utilised. The three-dimensional task has the format of a puzzle with three vertical pegs and a number of disks with a hole in the middle. The disks vary in size and can be placed around a peg. In the starting position of the task, all disks are arranged in pyramid form on one of the pegs, with the largest disk below and the smallest one on top. Respondents are asked to transfer the pyramid from the starting point to one of the other pegs. The third peg can also be used as a utility peg, but task solution behaviour is restricted by the rules that a) a disk can only be positioned on top of a disk that is larger, and b) putting a disk aside or c) moving more than one disk at the same time is not permitted (Noyes & Garland, 2003).

Dynamic Tower of Hanoi: pre-test and post-test

'Tower puzzles' with 2–7 disks were constructed for the pre-and post-test of the dynamic test 'Tower-puzzles'. Increasing or decreasing the number of disks in a Tower of Hanoi task has a considerable influence on task difficulty level. Although the standard version of the Tower of Hanoi uses uncoloured, wooden disks on one peg with the largest disk at the basis, in the current study Tower puzzles had coloured disks and starting positions with disks on one or two of the pegs. This modification enabled us to develop more fine-grained differences in the difficulty levels. Figure 2 presents some example Tower tasks that we utilised, involving different difficulty levels.

Both pre- and post-tests consisted of 12 puzzles with varying difficulty levels. As we considered it unlikely that participants would remember all the steps necessary to solve all pre-test puzzles, identical pre-and post-tests tasks were used. Testing started with an explanation of the task and solving rules, including an example puzzle with only two disks. All instructions were provided by the robot. The upper part of Appendix displays an overview of the progression in difficulty level of all 12 puzzles of the pre-and post-test, and provides the minimum number of steps required to solve each Tower puzzle, including the maximum number of steps permitted and the time limit for producing a solution. The pre- or post-test ended when three puzzles in a row were solved inaccurately or were unfinished because it was assumed that a child would then not be able to accurately solve subsequent tasks with higher difficulty levels. Such challenges could demotivate the child's eagerness to work with the robot. A response to a puzzle was scored as inaccurate when either the maximum of steps was reached, or the child had passed the time limit. The maximum number of steps was

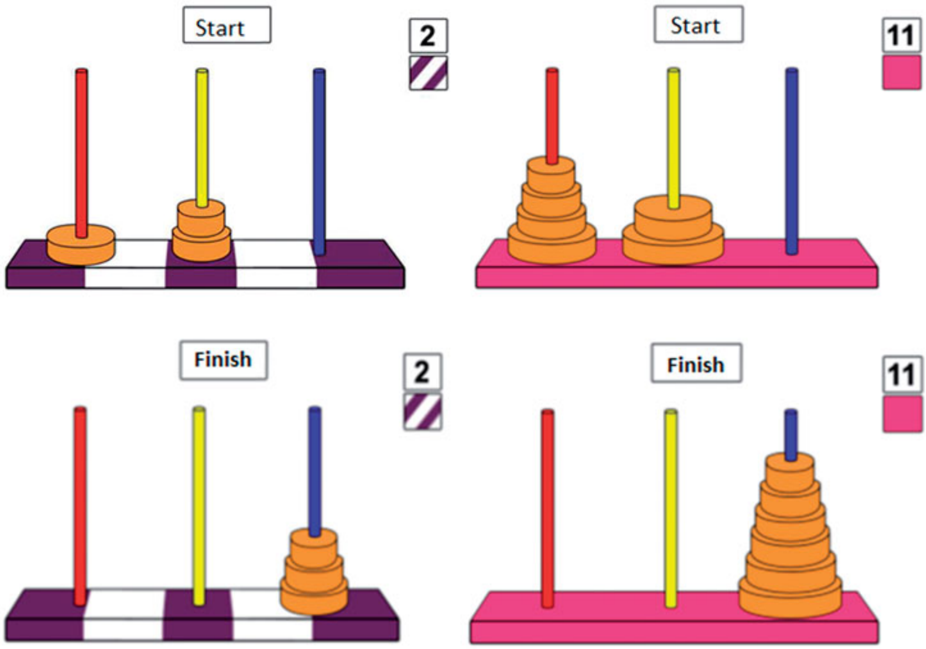

Figure 2. An easy and a difficult puzzle of the dynamic TOH test (at the start and finish).

calculated on the number of steps necessary for solving the task, including the possibility of starting with the wrong disk.

Dynamic Tower of Hanoi: training

The Tower of Hanoi training included a maximum of 16 Tower puzzles. The step-wise procedure provided graduated prompts starting with general, metacognitive prompts such as focussing and asking them to try to remember what they did before. If necessary, a child received cognitive prompts, including prompts regarding how to get a Tower puzzle onto the last peg. The last prompt included step-by-step scaffolds that were aimed at guiding the child to the solution of the puzzle. Prompts were provided when a child moved more than two consecutive steps away from the ideal solution track. Children received a maximum of five prompts per Tower puzzle. During training, children received a warning if they did not follow the Tower of Hanoi's solving rules: if they positioned a larger disk on a smaller one, placed more than one disk on the pegs at the same time, or placed a disk aside. In the training phase, prompts were distinguished from warnings, as the former focus on facilitating accurate independent task-solving, disclosing a small aspect of the task solving process, and the latter focus on task solving behaviours that are not permitted according to the Tower of Hanoi rules.

The children started the Tower of Hanoi training with an easy Tower puzzle including only 2 disks on one peg in the starting position. If they solved this first puzzle accurately without any prompts being provided, the Tower puzzles with respectively 3, 4, 5, 6, and 7 disks were administered. However, when one of the puzzles was

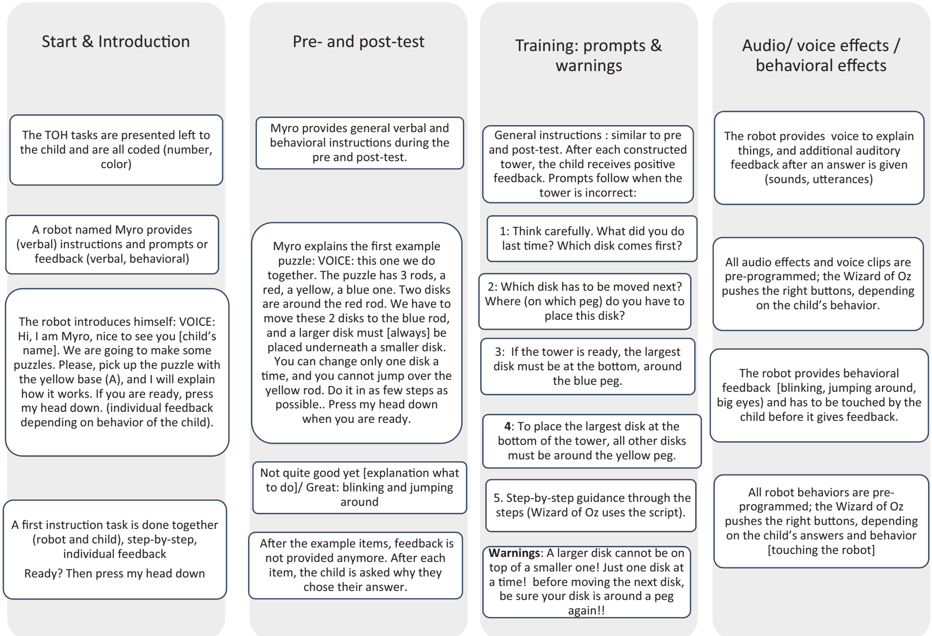

Figure 3. Schematic global overview of the instructions and feedback provided by the robot.

solved incorrectly, or took too many steps or too much solving time, an easier task was administered, with more disks on the third peg but with the same number of disks as the task they had previously solved. The children were then provided with prompts until the task was accurately solved. The training ended when the child, despite all prompts provided, still responded incorrectly to at least four of the last five Tower puzzles presented. The lower part of Appendix includes the order of training puzzles in schematic form. For all Tower puzzles, pre-programmed task instruction procedures were developed, based on task analyses of the Tower of Hanoi. On the basis of these analyses, the robot was programmed to interact with the child and give the necessary feedback (e.g. Noyes & Garland, 2003; Resing & Elliott, 2011). Figure 3 presents a schematic overview of the instructions and feedback provided by the robot.

Procedure

The test sessions took place once a week, over a period of three weeks. All children were seen individually at a quiet location at their school. The pre- and post-test tasks took approximately 30 minutes, and the training sessions about 45 minutes to administer. During all sessions, the robot interacted with the child, by giving feedback (voice, sounds, eyes, movements) and prompts following a standardised graduated prompts protocol (see Figure 3). The children were given a booklet containing on each page the start and end position of a Tower task. Every Tower had a different colour at the base and a number, whereby a child could be instructed which Tower they had to solve (see Figure 2 for two example pages of the booklet). The robot introduced himself to the child, said 'hello', named the child in person, and then started to explain the task, encouraging the child to solve the example Tower of Hanoi puzzle with him to see whether the task was understood. The child was then asked to pick a particular Tower task, and testing started, orally guided by the robot. Children had to touch the head of the robot as soon as they finished a task. Every step that was taken by child (and the robot) was saved in a log file.

Scoring and analyses

The Tower of Hanoi outcome variables analysed in this study were the number of accurately solved Tower puzzles, the total number of steps a child needed to solve the puzzles, the task completion time, the time taken before the child began to take physical action, the number of prompts needed during training, and the number of warnings during pre- and post-test stages.

Accuracy

For the accuracy variable, children's number of correct solutions (maximum = 12) within the time limit were counted. A solution was considered correct if, within the time limit, a child solved a puzzle with the minimum number of steps needed to solve the puzzle.

Task solving steps

The task solving step variable was defined as the number of solving steps a child needed to solve the puzzle correctly, within the time limit. If a child exceeded the maximum number of steps permitted when solving a puzzle, his or her answer would be scored as the maximum number of steps permitted for this particular puzzle (see Appendix). As the pre- or post-test ended when three consecutive puzzles were solved inaccurately within the allowed time frame, the number of steps children needed to solve the puzzles were defined by the actual number of steps for the puzzles administered, as well as the maximum number of steps of the puzzles they were not provided with.

Completion time

Completion time was defined as the mean completion time divided by the number of puzzles administered to the child.

Pre-solving time

Pre-solving (i.e. before a physical response commenced) time was based on the time between the start of the solving period (i.e. tower booklet and tower puzzle ready on the table) and the first step of the child. Pre-solving times of all the puzzles the child solved during pre-test, training, and post-test, and, per test session, were divided by the number of puzzles provided to the child. The resulting variable was the 'mean pre-solving time' per test session (pre-test, training, and post-test).

Prompts
The prompts variable consisted of the total number of prompts children received during training.

Warnings
Warnings were given if the child performed an action that was violated the solving rules of the Tower of Hanoi puzzle: each time a child positioned a larger disk on top of a smaller one, placed more than one disk on the pegs at the same time, or placed a disk aside.

Data analysis
Data were analysed with Pearson correlations, ANOVAs, repeated measures ANOVAs, and moderation analyses, using Hayes' (2018) PROCESS macro.

Results

Before analysing the data in relation to the research questions, two one-way analyses of variance (ANOVAs) were conducted to examine possible differences in age or initial level of inductive reasoning ability between children in the trained and non-trained condition. The analysis regarding age did not reveal significant differences between the children in the two conditions, F (1,36) = 0.988; p = .327. The analysis with level of inductive reasoning ability (Raven) as the dependent variable also did not show significant differences between the two groups of children, F (1,36) = 0.001; p = .974. Test-retest reliability was calculated by means of Pearson product-moment correlations and was found to be r = 0.618, p = .004 for the untrained children, and r = 0.347, p = .173 for the trained children, providing a preliminary indication of the validity of the dynamic test.

Effects of training on task-solving outcomes

Firstly, we examined progression overtime on the Tower of Hanoi tasks induced by repeated assessment, and the expected additional contribution of training to this progression. The behavioural outcomes we sought to measure were accuracy of task-solving, number of task-solving steps necessary to solve the Tower of Hanoi tasks pieces correctly, completion time, and pre-solving time. Table 2 provides the mean and standard deviations of the various dependent variables derived from dynamic testing.

Accuracy of task-solving
Changes in task-solving accuracy over time were examined with one within (Sessions: test sessions 1–2) and one between (Condition: training – no training) repeated measures ANOVA, with the number of accurately solved puzzles as the dependent variable. Significant effects for both sessions, F (1, 36) = 7.41, p = .010, $\eta_p^2 = 0.17$, and the interaction between Sessions and Condition, $F(1, 36) = 5.51$, p = .025, $\eta_p^2 = 0.14$; (sphericity assumed), were found. These outcomes, depicted in Figure 4a, reveal that both groups of children significantly increased their problem-solving accuracy between test

Table 2. Mean scores (M) and standard deviations (SD) per condition at pre-test and post-test sessions, for the dependent variables Accuracy, Number of steps, Completion time, and Pre-Solving time (seconds).

	Control group				Trained group			
	Pre-test		Post-test		Pre-test		Post-test	
	M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
Accuracy	5.20	1.70	5.30	1.69	5.24	1.79	6.59	1.18
Number of moves	483.95	21.56	490.05	18.92	500.94	20.28	472.12	18.31
Completion-time	101.12	29.37	69.33	12.44	86.98	18.49	76.91	18.45
Pre-solving time	12.84	3.36	14.24	4.49	13.12	2.67	14.68	5.32

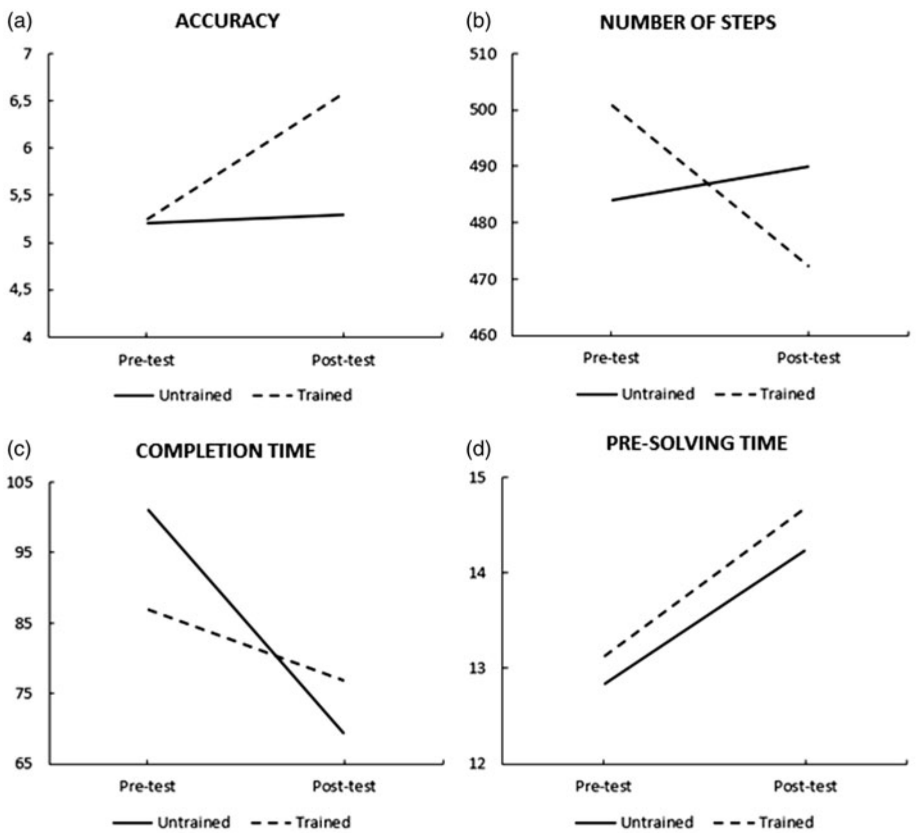

Figure 4. Effects of training on the following task-solving outcomes: (a) total number of accurately solved puzzles (accuracy), (b) total number of steps, (c) total completion time in seconds, (d) and mean pre-solving time in seconds.

sessions and trained children showed significantly more progression in task-accuracy than non-trained children.

Number of task-solving steps

Repeated measures ANOVA, with the dependent variable being the number of task-solving steps, Sessions (test sessions 1–2) as within, and Condition (training–non-training) as between, factors, also showed significant effects for Sessions and the interaction between Sessions and Condition, F (1, 36) = 9.08, p = .005, η_p^2 = 0.21; and F (1, 36) = 21.46, p = <.001, η_p^2 = .38 (sphericity assumed), respectively. Inspection of Figure 4b reveals that trained children, as expected, showed a significant decrease in the number of task-solving steps they needed to finish the tasks, while non-trained children failed to show such a decrease.

Task-completion time

Additional repeated measures ANOVA was performed with the mean task-completion time for each puzzle as the dependent variable. Further included were Sessions (test sessions 1-2) as within, and Condition (training – no training) as between factors. The outcomes of this analysis showed a significant effect for Sessions, F(1,36) = 28.32; p<.001, η_p^2 = 0.45, and a significant interaction between Sessions and Condition, F(1,36) = 7.61; p = .009, η_p^2 = 0.18. Inspection of Figure 4c showed that children in both condition groups reduced the time they needed for task-completion across sessions, and that, as partially expected, trained children took significantly more time to solve the tasks than the untrained children in the control group.

Pre-solving time

Repeated measures ANOVA with the mean pre-solving time over solved puzzles as the dependent variable, and Sessions (test sessions 1–2) as within, and Condition (training – no training) as between factors revealed a significant Sessions effect F(1,36) = 4.67; p = .038, η_p^2 = 0.12, but, in contrast to our expectation, there was no significant interaction between Sessions and Condition F(1,36) = 0.14; p = .905, η_p^2<0.01. Both trained and untrained children took significantly more pre-solving time during the post-test, as depicted in Figure 4d.

Number of prompts and warnings needed

The training on how to solve the Tower of Hanoi problems appeared to be rather difficult for the children, and revealed large individual differences between them in respect of the numbers of prompts and warnings needed. On average, children solved 3.18 Tower puzzles correctly (SD = 0.95) during the training. They required between 15 and 171 prompts, with a mean of 67.88 prompts (SD = 45.58).

Children were provided with a warning by the robot if they showed incorrect solving behaviour. The mean number of warnings was 2.8 (SD = 3.88; ranging from 0 to 15 times). These figures indicate that the children had to be given a considerable number of prompts during their training, but also showed large individual differences in this respect. The number of warnings given at pre-test (M_{pre} = 5.63 for experimental group; M_{pre} = 4.45 for control group children) decreased considerably for both groups of children (M_{post} = 1.75 for experimental and M_{post} = 0.55 for control group children). Repeated measures ANOVA with Sessions as within, and Condition as between, factors, and number of warnings as dependent variable, revealed a significant effect of Session, F(1,35) = 38.58; p<.001; η_p^2 = 0.53, but no significant interaction between Sessions and Condition F(1,35)<.01; p = 0.984, η_p^2 < 0.01.

We also considered to what extent the number of prompts required related to the number of warnings children received during pre-test, training, and post-test (see Table 3 for the correlation coefficients). In general, Pearson product-moment correlations revealed that the number of prompts correlated moderately and positively with the number of warnings children received during pre-test ($r = 0.329$, $p = .198$) , training ($r = 0.241$, $p = .578$), but modestly and negatively with the warnings received during post-test ($r = -0.358$, $p = .173$).

Planning and general executive functioning

It was considered that children's planning ability and executive functioning, in general, were likely to influence the progression in task solving accuracy from pre-test to post-test differently for dynamically tested children when compared with non-trained control group children. To check this, correlations between pre- and post-test scores and teachers' estimations of children's planning and executive function skills, plus two moderation analyses were conducted. The correlations are located in Table 3.

The first moderator analysis with the post-test accuracy scores on the Tower of Hanoi as the dependent variable, and pre-test accuracy scores, as the predictor, and Condition and the Planning and Organisation subscale of the Behaviour Rating Inventory of Executive Function as moderators revealed a significant Model ($p = .002$). Condition (p=.004), and Pre-test accuracy ($p = .001$) appeared to be the only significant predictor of the post-test scores of the children. Planning and various interactions did not reach statistical significance, as is revealed in Table 4

Comparable effects were shown in the second moderator analysis, with the post-test accuracy scores on the Tower of Hanoi as the dependent variable, and pre-test accuracy scores as the predicting variable, with Condition and the general executive functioning factor of the Behaviour Rating Inventory of Executive Function as moderators. Again, a significant Model ($p = .002$) was revealed, with Condition ($p = .004$) and Pre-test accuracy ($p = .001$) as significant moderators of the post-test scores of the children. The executive functioning and interaction variables did not reach statistical significance (see Table 5).

Exploring observations regarding the assessment procedure

The examiners who were involved in gathering the dynamic test outcomes also observed children's and classroom teachers' behaviour during the assessment period. They reported that, from the outset of dynamic testing with the tutoring peer robot, children were highly excited and motivated to work and 'play' with this robot, which was programmed to be as personable as possible and to know every child by name. The examiners observed that the children, across conditions, appreciated the testing sessions with the robot very much, and appeared highly motivated to solve the, sometimes very difficult, tasks together with him. In a very short time, they were talking to and about 'Myro' as if it were a patient teaching assistant, and hardly spoke to the examiner who was quietly sitting in the room at the computer screen. Pre-programmed instructions (pre and post-test) and the training procedure provided by the

Table 3. Correlations between the pre-test and post-test accuracy scores, and the BRIEF Planning and Organisation subscale and General Executive Functioning Index.

	Experimental condition							Control condition				
	Pre-test warn.	Post-test acc.	Post-test warn.	Prompts	Training warn.	Plan. & Organ.	Gen. Exec. Func.	Pre-test warn.	Post-test acc.	Post-test warn.	Plan. & Organ.	Gen. Exec. Func.
Pre-test acc.	0.188	0.347	−0.084	0.190	0.241	0.171	0.130	0.022	0.618**	−0.054	0.046	−0.135
Pre-test warn.	–	0.054	0.291	0.329	0.456	0.303	0.359	–	−0.408	−0.075	0.485*	0.428
Post-test acc.		–	−0.175	0.366	0.065	0.150	0.066		–	−0.104	−0.244	−0.375
Post-test warn.			–	−0.358	0.592*	0.191	0.306			–	0.191	0.181
Prompts				–	0.145	0.135	0.144				–	–
Training warn.					–	0.624**	0.726**					–
Plan. & Organ.						–	0.950**					0.834**

Note. $*p < .05$, $**p < .01$.

Table 4. Moderation analysis of the post-test accuracy scores predicted by planning and organisation, the pre-test and condition.

	B	*SE B*	*t*	*p*
Constant	5.920	0.212	27.880	<.0001
Pre-test	0.444	0.125	30.558	.001
Planning and organisation	−0.045	0.055	−0.817	.420
Pre-test × planning and organisation	−0.036	0.043	−0.812	.423
Condition	1.330	0.424	3.133	.004
Pre-test × condition	−0.310	0.259	−1.197	.241

Note. $R^2 = 0.447$.

Table 5. Moderation analysis of the post-test accuracy scores predicted by general executive functioning, the pre-test and condition.

	B	*SE B*	*t*	*p*
Constant	5.897	0.211	27.906	<.0001
Pre-test	0.448	0.127	3.526	.001
General executive functioning	−0.007	0.008	−0.885	.383
Pre-test × general executive functioning	−0.003	0.006	−0.459	.650
Condition	1.330	0.430	3.090	.004
Pre-test × condition	−0.355	0.252	−1.411	.168

Note. $R^2 = 0.441$.

robot were highly structured, and children often reacted to that in a (semi)interactive way, saying something like ‘I know that already, you told me before, Myro’.

Discussion

The present study sought to examine the effects of dynamic testing utilising complex problem solving tasks (e.g. Klahr & Robinson, 1981; Unterrainer et al., 2004). Instead of a human examiner, a pre-programmed table-top robot served as a training tool for 8-year-old children. The study showed that children’s task-solving accuracy generally improved when they were tested for a second time and that children’s progression in task accuracy varied considerably, depending on whether or not children were trained in solving Tower puzzles by the robot. Trained children not only showed significantly a greater progression in the number of Tower problems that they could solve accurately, they also used considerably fewer steps, although the Tower puzzles children were provided with increased rather quickly in difficulty level. These study outcomes reveal that the successful principles of dynamic testing with graduated prompts techniques, often applied in the field of inductive reasoning (e.g. Campione & Brown, 1987; Freund & Holling, 2011; Resing & Elliott, 2011; Stevenson, Heiser, & Resing, 2013; Tzuriel & George, 2009; Vogelaar et al., 2017) can be generalised to complex problems solving tasks as well, and that dynamic testing with Tower of Hanoi tasks lead to comparable, positive outcomes.

When children’s solving times are considered, two aspects are of particular importance. Firstly, both trained and untrained children took less time on the second testing phase. The mean completion time of trained children, however, decreased at a much slower rate than that of the untrained children. Secondly, both groups of children took considerably more time to think and plan ahead before they started the solving process. These two outcomes regarding solving time in combination with the differences in accuracy lead us to conclude that, most probably, both groups of children started to plan their solving steps more in advance, when they were retested, but did so in a rather different way. It must be noted that the Tower puzzles proved rather difficult for the 8-year old children involved in this study. At pre-test, children accurately completed between four and six of the tower tasks that increased in complexity rapidly. A study with older children may provide greater insight into the relationship between completion time, pre-solving time, and accuracy, because these relationships might not have a linear character, and might be moderated by task difficulty (e.g. Goldhammer et al., 2014).

The graduated prompts principles behind the training given by the robot were specifically designed to tap into each child's zone of proximal development (Serholt & Barendregt, 2016; Vygotsky, 1978; Wood, Bruner, & Ross, 1976). Inspection of the variation in the children's progression in relation to the outcomes revealed large individual differences. Of course, the potential extra value of these individualised outcomes of dynamic testing using approaches such as that outlined here will need to be further established and evaluated in future studies.

According to many scholars, planning is an important prerequisite for tasks involving complex reasoning, because it is necessary to think ahead and anticipate each of several problem-solving steps and consider appropriate use of strategy (e.g. Carlson et al., 2004; Lehto et al., 2003; Meltzer, 2018; Schiff & Vakil, 2015; Welsh, 1991; Welsh & Huizinga, 2005). Contrary to our expectations, planning skills, as judged by the children's teachers, did not moderate the accuracy scores of trained and untrained children. This suggests that children with poor planning skills may need more and shorter training sessions that utilise easier puzzles. It is also possible that the training provided to the children was insufficiently explicit in relation to the planning techniques. Comparing the study outcomes regarding completion and pre-solving time, and the role of planning, several additional explanations are worth taking into consideration. Firstly, it might be possible that the teachers' questionnaire used to measure the planning skills of healthy young children is not sensitive enough, as it was originally developed for children with problems in executive skills and employed as an indirect measure of teacher judgements regarding general executive activity (Toplak, West, & Stanovich, 2013). Secondly, pre-solving time, independently of whether children were trained, increased significantly when children were tested twice. The observed increase in pre-solving time may indicate that the children were encoding the task elements more thoroughly, thinking before doing, and breaking up the task in solvable steps (e.g. Alibali, Phillips, & Fischer, 2009). Children showed large individual differences in their pre-solving time. These individual differences may reflect differences in learning preferences, or task approach, but may also be indicative of the extent to which children solve a task analytically (Resing et al., 2012). In turn, this might provide an indication of the extent to which children plan their solution prior to starting the solving process. Of course, research is needed to further investigate this hypothesis.

Finally, children typically reach a successful level of Tower of Hanoi performance at around the ages 11–13 years (e.g. Bishop et al., 2001), and have not usually developed full planning skills at the age of 8. Future research may benefit from the deployment of a variety of instruments measuring planning with children of different ages. Moreover, complex problem-solving tasks, such as the Tower of Hanoi, have been said to suffer from task impurity, which means that solving the task requires or taps into different aspects of executive functions and non-executive cognitive processes (Packwood, Hodgetts, & Tremblay, 2011). In the case of the Tower of Hanoi often mentioned executive functioning aspects measured include (visual-spatial) working memory and inhibition (Welsh, Satterlee-Cartmell, & Stine, 1999).

Unlike earlier study outcomes with computerised forms of dynamic testing (e.g. Resing et al., 2012, 2017; Tzuriel & Shamir, 2002), and with virtual reality tasks (Passig et al., 2016), the present study focussed on the utility of computerised dynamic testing (with the robot as support examiner) in combination with the manipulation of 3 D task tangibles. It is particularly the case that for younger persons, computerised versions of the Tower of Hanoi task are more difficult than its normal, 3D representation (e.g. Schiff & Vakil, 2015). Inspection of our findings, however, showed that our computerised table-top robot generated considerable progression in problem-solving performance. The merits of using a robot as a support assistant in dynamic testing seem obvious. Earlier studies (Henning, Verhaegh, & Resing, 2010; Resing et al., 2017; Resing & Elliott, 2011; Veerbeek, Vogelaar, Verhaegh, & Resing, 2019) have already shown the positive aspects of the use of an electronic console for dynamic testing. The current study replicates this but within a different task domain, and goes one step further. With the introduction of the robot, the children could, in a seamless learning setting, freely move the disks in a three-dimensional space (e.g. Fößl, Ebner, Schön, & Holzinger, 2016; O'Malley & Stanton Fraser, 2004; Schmitz, Klemke, Walhout, & Specht, 2015). Additionally, the robot proved to be an active, enjoyable companion, with both verbal and non-verbal interaction qualities, with – for the moment – the examiner as a quasi-Wizard of Oz figure in the background. Further studies, building upon this work, could assist in understanding how we can best assess children's potential for learning (Clabaugh, Ragusa, Sha, & Matarić, 2015; Granott, 2005). An important consideration of such studies should be to what extent the study findings can be generalised to other contexts, taking into account factors such as children's age and cultural background. Perhaps in older children or children with a different cultural background, or less or more exposure to technology in education, different results would be obtained. Although the Wizard of Oz setting utilised in the current study enabled objective dynamic testing in combination with working with authentic materials, having a human examiner functioning as the eyes and ears of the robot would be highly labour-intensive, and thus expensive, if the test is to be used for a large number of children. Therefore, in future studies it should be investigated to what extent the robot could be programmed utilising strict if-then protocols on the basis of children's solutions that human interference is no longer necessary.

Although the robot had a number of limitations, such as repeating instructions using the same words repeatedly, the children were highly responsive and motivated to work with it and even provided the robot with feedback. In future studies, it would be interesting to examine whether children in the "control" condition to some extent also learned by doing from the assessment by the robot as they were motivated for the testing sessions. An extension of the study design with one or more extra control groups, for instance, a control group that practices solving puzzles but does not receive help or feedback between pre-test and post-test, and extra variables, such as a different training procedure and training by human versus training by robot, would seem valuable. The robot provided prompts to the child when needed, but these were not yet always optimally tailored to the particular problem-solving procedures that the children sometimes demonstrated. Further research could be targeted to the development of highly adaptive and differentiated interaction responses using prompts that are programmed into a computerised robot. In relation to the complex problem-solving domain studied here, future research might focus on the further fine-tuning of prompts, and dynamic scaffolds, appropriate methods to encourage planning skills and reduce cognitive load, and consideration of idiosyncratic approaches children showed when tackling the items (e.g. Granott, 2005, Khandelwal & Mazalek, 2007).

As noted above, the amount of valuable data that can be obtained during a dynamic testing session is far greater than can be recorded contemporaneously by paper and pencil. We anticipated, and found that the robot technology could greatly assist us in assessing and examining the problem-solving processes taking place during dynamic testing – a key aspect of process-oriented dynamic testing (Elliott et al., 2010; Jeltova et al., 2011; Resing et al., 2017; Sternberg & Grigorenko, 2002). As other empirical studies reporting the effects of robots as instructional or assessment tools involve learning in one way or the other, the findings in the current study should provide further opportunities and inspiration for research within the broader field of learning and training complex cognitive skills (e.g. Benitti, 2012).

We are aware that considerable development of both hardware and software will be necessary before small table robots are fully capable of supporting educational psychologists and teachers in undertaking complex assessment of children's cognitive current and potential functioning (e.g. Timms, 2016). We believe, however, that the results of our study demonstrate that even a simplified prototype version of such a robot, with its instructive teaching possibilities, friendly appearance, and patience, can stimulate and motivate children in learning how to solve complex cognitive tasks within an authentic context. Meaningful engagement, and the provision of detailed feedback to teachers on the child's problem-solving trajectory, should subsequently result in more closely tailored instruction that, in combination, should ultimately, have an important impact upon the development of children's cognitive and academic growth (e.g. Jones, Bull, & Castellano, 2018; Mubin et al., 2013; Wood, 2001).

Acknowledgements

We would like to thank Bart Dirkx and Ruud van der Aalst for their help in the construction phase and for letting us use their robot in development; Floor Thissen for her help with the manuscript; and Nathalie IJsselmuiden, Huguette Fles, and Karlijn Nigg for their help in collecting the data for this study.

Disclosure statement

No potential conflict of interest was reported by the authors.

ORCID

Wilma C. M. Resing http://orcid.org/0000-0003-3864-4517
Bart Vogelaar http://orcid.org/0000-0002-5131-2480
Julian G. Elliott http://orcid.org/0000-0002-9165-5875

References

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

André, V., Jost, C., Hausberger, M., Le Pévédic, B., Jubin, R., Duhaut, D., & Lemasson, A. (2014). Ethorobotics applied to human behaviour: Can animated objects influence children's behaviour in cognitive tasks? *Animal Behaviour*, *96*, 69–77. doi:10.1016/j.anbehav.2014.07.020

Baxter, P., Ashurst, E., Read, R., Kennedy, J., & Belpaeme, T. (2017). Robot education peers in a situated primary school study: Personalisation promotes child learning. *PLoS One*, *12*, e0178126. doi:10.1371/journal.pone.0178126

Belpaeme, T., Kennedy, J., Ramachandran, A., Scassellati, B., & Tanaka, F. (2018). Social robots for education: A review. *Science Robotics*, *3*(21). doi:10.1126/scirobotics.aat5954

Benitti, F. B. V. (2012). Exploring the educational potential of robotics in schools: A systematic review. *Computers and Education*, *58*, 978–988. doi:10.1016/j.compedu.2011.10.006

Beran, T., & Ramirez-Serrano, A. (2010). Do children perceive robots as alive? Children's contributions of human characteristics. In *Proceedings of the 5th ACM/IEEE international conference on Human-robot interaction* (pp. 137–138). Piscataway, NJ: IEEE Press.

Berg, C. A., Strough, J., Calderone, K., Meegan, S. P., & Sansone, C. (1997). Planning to prevent everyday problems from occurring. In S. L. Friedman & E. K. Scholnick (Eds.), *The developmental psychology of planning: Why, how, and when do we plan?* (pp. 209–236). Mahwah, NJ: Lawrence Erlbaum Associates Publishers.

Bishop, D. V. M., Aamodt-Leeper, G., Creswell, C., McGurk, R., & Skuse, D. H. (2001). Individual differences in cognitive planning on the Tower of Hanoi task: Neuropsychological maturity or measurement error? *Journal of Child Psychology and Psychiatry*, *42*(4), 551–556. doi:10.1111/1469-7610.00749

Brown, L., & Howard, A. M. (2013). *Engaging children in math education using a socially interactive humanoid robot*. Paper presented at the 13th IEEE-RAS International Conference on Humanoid Robots, Atlanta, GA: IEEE.

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, NY: Guilford Press.

Carlson, S. M., Moses, L. J., & Claxton, L. J. (2004). Individual differences in executive functioning and theory of mind: An investigation of inhibitory control and planning ability. *Journal of Experimental Child Psychology*, *87*, 299–319. doi:10.1016/j.jecp.2004.01.002

Cha, E., Greczek, J., Song, A., & Matarić, M. J. (2017). My classroom robot: Exploring telepresence for K-12 education in a virtual environment. In *26th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)* (pp. 689–695). Piscataway, NJ: IEEE Press.

Chang, C.-W., Lee, J.-H., Chao, P.-Y., Wang, C.-Y., & Chen, G.-D. (2010). Exploring the possibility of using humanoid robots as instructional tools for teaching a second language in primary school. *Educational Technology and Society*, *13*, 13–24.

Chin, K.-Y., Hong, Z.-W., & Chen, Y.-L. (2014). Impact of using an educational robot-based learning system on students' motivation in elementary education. *IEEE Transactions on Learning Technologies*, *7*, 333–345. doi:10.1109/TLT.2014.2346756

Clabaugh, C., Ragusa, G., Sha, F., & Matarić, M. (2015). *Designing a socially assistive robot for personalized number concepts learning in preschool children*. Paper presented at the 2015 Joint

IEEE International Conference on Development and Learning and Epigenetic Robotics. Providence, RI: IEEE.

Culbertson, W. C., & Zillmer, E. A. (1998). The Tower of London DX: A standardized approach to assessing executive functioning in children. *Archives of Clinical Neuropsychology*, *13*, 285–301. doi:10.1093/arclin/13.3.285

Dahlbäck, N., Jönsson, A., & Ahrenberg, L. (1993). Wizard of Oz studies – why and how. *Knowledge-Based Systems*, *6*, 258–266. doi:10.1016/0950-7051(93)90017-N

Deublein, A. (2018). Scaffolding of motivation in learning using a social robot. *Computers and Education*, *125*, 182–190. doi:10.1016/j.compedu.2018.06.015

Elliott, J. G., Grigorenko, E. L., & Resing, W. C. M. (2010). Dynamic assessment: The need for a dynamic approach. In P. Peterson, E. Baker, & B. McGaw (Eds.), *International Encyclopedia of Education* (Vol. 3, pp. 220–225). Amsterdam, The Netherlands: 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. doi:10.1080/00131911.2018.1396806

Fößl, T., Ebner, M., Schön, S., & Holzinger, A. (2016). A field study of a video supported seamless-learning-setting with elementary learners. *Educational Technology and Society*, *19*, 321–336.

Freund, P. A., & Holling, H. (2011). How to get really smart: Modeling retest and training effects in ability testing using computer-generated figural matrix items. *Intelligence*, *39*, 233–243. doi:10.1016/j.intell.2011.02.009

Gioia, G. A., Isquith, P. K., Guy, S. C., & Kenworthy, L. (2000). Behavior Rating Inventory of Executive Function. *Child Neuropsychology*, *6*, 235–238. doi:10.1076/chin.6.3.235.3152

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*, 608–626. doi:10.1037/a0034716

Goswami, U. (2013). The development of reasoning by analogy. In P. Barrouillet & C. Gauffroy (Eds.), *The development of thinking and reasoning* (pp. 49–70). London, UK: Psychology Press.

Granott, N. (2005). Scaffolding dynamically toward change: Previous and new perspectives. *New Ideas in Psychology*, *23*, 140–151. doi:10.1016/j.newideapsych.2006.07.002

Grigorenko, E. L. (2009). Dynamic assessment and response to intervention: Two sides of one coin. *Journal of Learning Disabilities*, *4*, 111–132. doi:10.1177/0022219408326207

Hayes, A. F. (2018). *Introduction to mediation, moderation, and conditional process analysis: A regression-based approach*. New York, NY: The Guilford 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 and Child Psychology*, *28*, 85–100.

Hinz, A. M., Klavžar, S., Milutinović, U., & Petr, C. (2018). *The Tower of Hanoi – Myths and Maths*. Cham, Switzerland: Birkhäuser.

Hong, Z.-W., Huang, Y.-M., Hsu, M., & Shen, W.-W. (2016). Authoring robot-assisted instructional materials for improving learning performance and motivation in EFL Classrooms. *Journal of Educational Technology and Society*, *19*, 337–349.

Huang, S.-H., Wu, T.-T., Chu, H.-C., & Hwang, G.-J. (2008). March). *A decision tree approach to conducting dynamic assessment in a context-aware ubiquitous learning environment*. Paper presented at the Fifth IEEE International Conference on Wireless, Mobile, and Ubiquitous Technology in Education, Beijing, China: IEEE.

Huizinga, M., & Smidts, D. P. (2010). Age-related changes in executive function: A normative study with the Dutch version of the Behavior Rating Inventory of executive function. *Child Neuropsychology*, *17*(1), 51–66. doi:10.1080/09297049.2010.509715

Hussain, S. L., Lindh, J., & Shukur, J. G. (2006). The effect of LEGO training on pupils' school performance in mathematics, problem solving ability and attitude: Swedish data. *Educational Technology and Society*, *9*, 182–194.

Jeltova, I., Birney, D., Fredine, N., Jarvin, L., Sternberg, R. J., & Grigorenko, E. L. (2011). Making instruction and assessment responsive to diverse students' progress: Group-administered dynamic assessment in teaching mathematics. *Journal of Learning Disabilities*, *44*, 381–395. doi:10.1177/0022219411407868

Jones, A., & Castellano, G. (2018). Adaptive robotic tutors that support self-regulated learning: A longer-term investigation with primary school children. *International Journal of Social Robotics*, *10*(3), 357–370. doi:10.1007/s12369-017-0458-z

Jones, A., Bull, S., & Castellano, G. (2018). I know that now, I'm going to learn this next" promoting self-regulated learning with a robotic tutor. *International Journal of Social Robotics*, *10*, 439–454. doi:10.1007/s12369-017-0430-y

Kanero, J., Geçkin, V., Oranç, C., Mamus, E., Küntay, A. C., & Göksun, T. (2018). Social robots for early language learning: Current evidence and future directions. *Child Development Perspectives*, *12*, 146–151. doi:10.1111/cdep.12277

Kaufman, S. B. (2007). Sex differences in mental rotation and spatial visualization ability: Can they be accounted for by differences in working memory capacity? *Intelligence*, *35*, 211–223. doi:10.1016/j.intell.2006.07.009

Khandelwal, M. (2006). *Teaching table: A tangible mentor for pre-kindergarten math education* (Unpublished master's thesis). Georgia Tech University, Atlanta, GA.

Khandelwal, M., & Mazalek, A. (2007). Teaching table. In *Proceedings of the 1st International Conference on Tangible and Embedded Interaction, Chapter 4, learning through physical interaction* (pp. 191–194). Baton Rouge, LA: ACM.

Klahr, D., & Robinson, M. (1981). Formal assessment of problem-solving and planning processes in preschool children. *Cognitive Psychology*, *13*(1), 113–148. doi:10.1016/0010-0285(81)90006-2

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

Kotovsky, K., Hayes, J. R., & Simon, H. A. (1985). Why are some problems hard? Evidence from Tower of Hanoi. *Cognitive Psychology*, *17*, 248–294. doi:10.1016/0010-0285(85)90009-X

Kozima, H., & Nakagawa, C. (2007). *A robot in a playroom with preschool children: Longitudinal field practice*. Paper presented at the 16th IEEE International Conference Robot and Human Interactive Communication, Jeju, South Korea: IEEE. doi:10.1109/ROMAN.2007.4415238

Lehto, J. E., Juujärvi, P., Kooistra, L., & Pulkkinen, L. (2003). Dimensions of executive functioning: Evidence from children. *British Journal of Developmental Psychology*, *21*(1), 59–80. doi:10.1348/026151003321164627

Libin, A. V., & Libin, E. V. (2010). Person-robot interactions from the robopsychologists' point of view: The robotic psychology and robotherapy approach. *Proceedings of IEEE*, 92, 1789–1803. doi:10.1109/JPROC.2004.835366

Meltzer, L. (2018). *Executive function in education: From theory to practice*. New York, NY: Guilford Press.

Moriguchi, Y., Kanda, T., Ishiguro, H., Shimada, Y., & Itakura, S. (2011). Can young children learn words from a robot? *Interaction Studies*, *12*(1), 107–118. doi:10.1075/is.12.1.04mor

Movellan, J. R., Eckhardt, M., Virnes, M., & Rodriguez, A. (2009). *Sociable robot improves toddler vocabulary skills*. Paper presented at the 4th ACM/IEEE International Conference on Human-Robot Interaction, La Jolla, CA: IEEE.

Mubin, O., Stevens, C. J., Shahid, S., Al Mahmud, A., & Dong, J. J. (2013). A review of the applicability of robots in education. *Journal of Technology in Education and Learning*, *1*, 1–7. doi:10.2316/journal.209.2013.1.209-0015

Noyes, J. M., & Garland, K. J. (2003). Solving the Tower of Hanoi: Does mode of presentation matter? *Computers in Human Behavior*, *19*, 579–592. doi:10.1016/S0747-5632(03)00002-5

O'Malley, C., & Stanton Fraser, D. (2004). Literature review in learning with tangible technologies. *FutureLab*. Retrieved from www.futurelab.org.uk/research/lit_reviews.htm

Packwood, S., Hodgetts, H. M., & Tremblay, S. (2011). A multiperspective approach to the conceptualization of executive functions. *Journal of Clinical and Experimental Neuropsychology*, *33*, 456–470. doi:10.1080/13803395.2010.533157

Passig, D., Tzuriel, D., & Eshel-Kedmi, G. (2016). Improving children's cognitive modifiability by dynamic assessment in 3D immersive virtual reality environments. *Computers and Education*, *95*, 296–308. doi:10.1016/j.compedu.2016.01.009

Raven, J., Raven, J. C., & Court, J. H. (2003). *Manual for Raven's progressive matrices and vocabulary scales. Section 1: General overview*. San Antonio, TX: Harcourt Assessment.

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

Resing, W. C. M., Bakker, M., Elliott, J. G., & Vogelaar, B. (2019). Dynamic testing: Can a robot as tutor be of help in assessing children's potential for learning? *Journal of Computer-Assisted Learning*, *35*, 540–554. doi:10.1111/jcal.1235814RESINGETAL

Resing, W. C. M., Steijn, W. M. P., Xenidou-Dervou, I., Stevenson, C. E., & Elliott, J. G. (2011). Computerized dynamic testing: A study of the potential of an approach using sensor technology. *Journal of Cognitive Education and Psychology*, *10*, 178–194. doi:10.1891/1945-8959.10.2.178

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*, 173–191. doi:10.1080/01443410.2016.1164300

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*, 231–251. doi:10.1891/1945-8959.14.2.231

Resing, W. C. M., Xenidou-Dervou, I., Steijn, W. M., & 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. doi:10.1016/j.lindif.2011.11.002

Salnaitis, C., Baker, C. A., Holland, J., & Welsh, M. (2011). Differentiation Tower of Hanoi performance: Interactive effects of psychopathic tendencies, impulsive response styles, and modality. *Applied Neuropsychology*, *18*(1), 37–46. doi:10.1080/09084282.2010.523381

Schiff, R., & Vakil, E. (2015). Age differences in cognitive skill learning, retention and transfer: The case of the Tower of Hanoi Puzzle. *Learning and Individual Differences*, *39*, 164–171. doi:10.1016/j.lindif.2015.03.010

Schmitz, B., Klemke, R., Walhout, J., & Specht, M. (2015). Attuning a mobile simulation game for school children using a design-based research approach. *Computers and Education*, *81*, 35–48. doi:10.1016/j.compedu.2014.09.001

Serholt, S., & Barendregt, W. (2016). October). *Robots tutoring children: Longitudinal evaluation of social engagement in child-robot interaction*. Paper presented at the 9th Nordic Conference on Human-Computer Interaction. Gothenburg, Sweden: ACM.

Serholt, S., Basedow, C. A., Barendregt, W., & Obaid, M. (2014). November). *Comparing a humanoid tutor to a human tutor delivering an instructional task to children*. Paper presented at the 14th IEEE-RAS International Conference on Humanoid Robots. Madrid, Spain: IEEE.

Simon, H. A. (1975). The functional equivalence of problem solving skills. *Cognitive Psychology*, *7*, 268–288. doi:10.1016/0010-0285(75)90012-2

Smidts, D. P., & Huizinga, M. (2009). *BRIEF executieve functies gedragsvragenlijst: Handleiding [BRIEF executive functions questionnaire: Manual]*. Amsterdam, The Netherlands: Hogrefe.

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

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*, 159–169. doi:10.1016/j.cedpsych.2013.02.001

Stevenson, C. E., Touw, K. W. J., & Resing, W. C. M. (2011). Computer or paper analogy puzzles: Does assessment mode influence young children's strategy progression? *Educational and Child Psychology*, *28*, 67–84.

Sullivan, F. R. (2008). Robotics and science literacy: Thinking skills, science process skills and systems understanding. *Journal of Research in Science Teaching*, *45*, 373–394. doi:10.1002/tea.20238

Tanaka, F., & Matsuzoe, S. (2012). Children teach a care-receiving robot to promote their learning: Field experiments in a classroom for vocabulary learning. *Journal of Human-Robot Interaction*, *1*, 78–95. doi:10.5898/JHRI.1.1.Tanaka

Tanaka, F., Cicourel, A., & Movellan, J. R. (2007). Socialization between toddlers and robots at an early childhood education center. *Proceedings of the National Academy of Sciences of Sciences*, *104*, 17954–17968. doi:10.1073/pnas.0707769104

Timms, M. J. (2016). Letting artificial intelligence in education out of the box: Educational robots and smart classrooms. *International Journal of Artificial Intelligence in Education*, *26*, 701–712. doi:10.1007/s40593-016-0095-y

Toh, L. P. E., Causo, A., Tzuo, P. W., Chen, I., & Yeo, S. H. (2016). A review on the use of robots in education and young children. *Journal of Educational Technology and Society*, *19*, 148.

Toplak, M. E., West, R. F., & Stanovich, K. E. (2013). Practitioner review: Do performance-based measures and ratings of executive function assess the same construct? *Journal of Child Psychology and Psychiatry*, *54*, 131–143. doi:10.1111/jcpp.12001

Tzuriel, D., & George, T. (2009). Improvement of analogical reasoning and academic achievement by the Analogical Reasoning Programme (ARP). *Educational and Child Psychology*, *26*, 71–93.

Tzuriel, D., & Shamir, A. (2002). The effects of mediation in computer assisted dynamic assessment. *Journal of Computer Assisted Learning*, *18*(1), 21–32. doi:10.1046/j.0266-4909.2001.00204.x

Tzuriel, D., Isman, E. B., Klung, T., & Haywood, H. C. (2017). Effects of teaching classification on classification, verbal conceptualization, and analogical reasoning in children with developmental language delays. *Journal of Cognitive Education and Psychology*, *16*(1), 107–124. doi:10.1891/1945-8959.16.1.107

Unterrainer, J. M., Rahm, B., Kaller, C. P., Leonhart, R., Quiske, K., Hoppe-Seyler, K., … Halsband, U. (2004). Planning abilities and the Tower of London: is this task measuring a discrete cognitive function? *Journal of Clinical and Experimental Neuropsychology*, *26*, 846–856. doi:10.1080/13803390490509574

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*, 260–280. doi:10.1891/1945-8959.16.3.260

Veerbeek, J., Vogelaar, B., Verhaegh, J., & Resing, W. C. M. (2019). Process-oriented measurement in dynamic testing using electronic tangibles. *Journal of Computer Assisted Learning*, *35*(1), 127–147. doi:10.1111/jcal.12318

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

Vogelaar, B., & Resing, W. C. M. (2018). Changes over time and transfer of analogy-problem solving of gifted and non-gifted children in a dynamic testing setting. *Educational Psychology*, *38*, 898–914. doi:10.1080/01443410.2017.1409886

Vogelaar, B., Bakker, M., Hoogeveen, L., & Resing, W. C. M. (2017). Dynamic testing of gifted and average-ability children's analogy problem solving: Does executive functioning play a role? *Psychology in the Schools*, *54*, 837–851. doi:10.1002/pits.22032

Vygotsky, L. S. (1978). *Mind in society: The development of higher psychological processes* (M. Cole, V. John-Steiner, S. Scribner, & E. Sonberman, Trans.). Cambridge, MA: Harvard University Press (original work published 1938).

Welsh, M. C. (1991). Rule-guided behavior and self-monitoring on the Tower of Hanoi disk-transfer task. *Cognitive Development*, *62*, 59–67. doi:10.1016/0885-2014(91)90006-Y

Welsh, M. C., & Huizinga, M. (2001). The development and preliminary validation of the Tower of Hanoi-Revised. *Assessment*, *8*, 167–176. doi:10.1177/107319110100800205

Welsh, M. C., & Huizinga, M. (2005). Tower of Hanoi disk-transfer task: Influences of strategy knowledge and learning on performance. *Learning and Individual Differences*, *15*, 283–298. doi:10.1016/j.lindif.2005.05.002

Welsh, M. C., Satterlee-Cartmell, T., & Stine, M. (1999). Towers of Hanoi and London: Contribution of working memory and inhibition to performance. *Brain and Cognition*, *41*, 231–242. doi:10.1006/brcg.1999.1123

WittyWorX. (2012). WittyWorX. Retrieved from http://www.wittyworx.com/index

Wood, D. (2001). Scaffolding, contingent tutoring, and computer-supported learning. *International Journal of Artificial Intelligence in Education*, *12*, 280–293.

Wood, D., Bruner, J. S., & Ross, G. (1976). The role of tutoring in problem solving. *Journal of Child Psychology and Psychiatry*, *17*, 89–100. doi:10.1111/j.1469-7610.1976.tb00381.x

Wu, H. M., Kuo, B. C., & Wang, S. C. (2017). Computerized dynamic adaptive tests with immediately individualized feedback for primary school mathematics learning. *Educational Technology and Society*, *20*, 61–72.

Zimmerman, B. J. (2002). Becoming a self-regulated learner: An overview. *Theory into Practice*, *41*, 64–70. doi:10.1207/s15430421tip4102_2

Appendix

Schematic overview of the puzzle-sequence during training

When a child solves all puzzles correctly, he or she will complete the following series of puzzles:

Table A1. Puzzle characteristics of the dynamic TOH pre- and post-test.

Puzzle	Number of disks	Minimum no. of steps	Maximum no. of steps	Time limit in minutes
1	3	4	19	5
2	3	7	22	5
3	4	9	24	5
4	4	11	26	5
5	4	13	28	5
6	4	15	30	5
7	5	20	32	6
8	5	25	37	6
9	5	31	43	6
10	6	47	57	7
11	6	63	76	7
12	7	127	153	8

2, 3, 4, 5, 6, 7. However, when a child provides an incorrect answer, he or she will continue along a different pathway. For example, when puzzle 4 is solved incorrectly, the child will continue to puzzle 4a, after which the child will continue to puzzle 4b if puzzle 4a was solved incorrectly, or to puzzle 5 if 4a was solved correctly. In Table 2, an overview of this system is provided. In cases where a child solves four out of five puzzles incorrectly, the training will be terminated after the next puzzle. For example, a child completes the following puzzles: puzzle 4 incorrectly, 4a incorrectly, 4b correctly, 5 incorrectly, and 5a incorrectly. In this case, the child will still be asked to solve puzzle 5b, after which the training will be terminated irrespective of whether the child manages to solve this puzzle correctly. A puzzle is categorised as incorrect, if the child needs a hint to solve it correctly.

Table A2. The training procedure and puzzle-sequence during training.

Puzzle number	If correct, go to	If not correct, go to	Number of steps required (minimum)
2	3	3	3
3	4	3a	7
3a	4	3b	5
3b	4	4	4
4	5	4a	15
4a	5	4b	13
4b	5	5	10
5	6	5a	31
5a	6	5b	25
5b	6	6	20
6	7	6a	63
6a	7	6b	45
6b	7	7	38
7	Stop	7a	127
7a	Stop	7b	105
7b	Stop	Stop	84