Logical reasoning in a deductive version of the Mastermind game

Student: Gabriela-Olivia Trușescu, s0996963

Master specialization: Educational Studies, Leiden University

Supervisor: Prof. dr. Maartje Raijmakers, Leiden University

Second assessor: Dr. Arnout Koornneef

Abstract

The aim of this study was to investigate whether individuals develop and adapt strategies with experience in a deductive reasoning task. 23 first-year students performed both a working memory- and a deductive reasoning task, while their eye movements were recorded by an eye-tracker. Repeated-measures MANCOVA revealed that students became more efficient, selective and accurate across blocks. They tended to process information from top to bottom, which indicates that order of processing information has a contribution in performing the task. Findings show that people develop more efficient strategies over time. Working memory plays a role in learning, good memorizers developing the most efficient solution strategies. I discuss findings in relation to a logical model and practical implications.

Keywords: deductive reasoning, selective strategy, working memory, learning, eye-tracking

Introduction and theoretical framework

The abilities to reason, to solve problems and make decisions in everyday life represent crucial aspects of human intelligence (Holyoack & Morrison, 2005). Logical reasoning is an instrument that helps individuals make sense of the world, by inferring or drawing conclusions about various phenomena, based on given information. The ability to understand a text, to identify patterns, to generalize information in order to produce more general rules or principles (Lohman & Lakin, 2009) is an important skill to acquire during school time. As a matter of fact, it is not surprising that the results of reasoning tests were found as predictor of academic success in 7th to 10th-grade students (Yenilmez, Sungur & Tekkaya, 2006; Vock, Preckel & Holling, 2011; Bhat, 2016).

The main theories that explain logical reasoning are mental rules and mental model theories, which could apply to different types of logical problems, such as syllogisms (Stanovich, Sá & West, 2004). The mental model theory (Johnson- Laird, 2004) states that first of all, individuals transform premises into a mental model that is in line with the premises. Not everyone is capable to build a coherent representation of the information from the premises: people with a poor working memory capacity have a limited ability to produce sufficient models to assess the validity of a conclusion (Lohman & Lakin, 2009). According to the theory of mental rules (Rips, 1994), logical reasoning in solving deductive problems involves three mental processes: translating the premises into representations stored in working memory, applying rule-based schemas to the representations to come to a conclusion and applying other rules to check the working memory content for inconsistencies (Lohman & Lakin, 2009). For each of these steps, there is a probability of making errors, which are attributed to working memory overload (Gilhooly, 2004). According to these theories, working memory plays a crucial role in the logical reasoning process. Limitations in the capacity to store information have an impact not only in creating a mental model of the task at hand, but also on the way people manipulate rules in order to find a solution.

In the current study, we aim to investigate the relationship between reasoning strategies, working memory and learning. Specifically, we intend to study the way strategies used by people improve with learning, and the impact of working memory on the learning process.

Strategy complexity, working memory and performance

Working memory is the ability to actively process and store information on a short-term basis (Cowan, 1995), and represents a key-factor in the reasoning process. It is usually measured by the operation span test, which assesses the number of items a subject can hold in the short-term memory, while manipulating further information (Copeland & Radvansky, 2004), and the extent to which they can cope with interference (Lustig, May & Hasher, 2001). An example of a typical working memory task requires first of all that participants comprehend and follow complex instructions, and then find and develop strategies to perform difficult tasks. Moreover, participants have to process simultaneously a number of items while remembering others. Working memory tasks demand great mental effort, especially when the memory load is high. There are complex cognitive processes involved in a working memory task, such as controlling the order of operations, suppressing incorrect responses, organizing information from various domains and integrating concepts into a coherent mental representation (Lohman & Lakin, 2009).

The ability to handle tasks of different complexity level varies individually. People with a good working memory capacity had also a higher reasoning ability and performed better in building multiple models, because they used more complex strategies (Gilhooly et al., 1999). Moreover, sometimes people with great spans tended to invent their own shortcut rules while progressing through the task, even though their methods were wrong. They started by applying a complex procedure, but in the meantime, found a shorter way, which helped them arrive faster at a solution. With the development of a simpler strategy, individuals decreased the cognitive load and answered quickly and effortlessly, that is, to use more effective strategies (Hinze, Bunting & Pellegrino, 2009). In contrast, people with a poor working memory showed a limited ability to produce sufficient models to assess the validity of a conclusion; they choose easier and incorrect procedures, as they found it difficult to deal with too sophisticated strategies. Those with a smaller span had troubles in finding a solution and were more likely to conclude based on the initial model. They preferred one-model syllogisms and needed more time to reach a conclusion, while struggling with building mental models (Copeland & Radvansky, 2004).

Empirical evidence shows that subjects applied different strategies in solving a problem, for a better speed and accuracy. Efficient solvers learned the rules fast and worked more independently, without asking for help. In contrast, less efficient subjects needed more time, overused help to complete the task and did not prove a consistent achievement. Previous research

showed that working memory was a key-aspect not only in performance, but in developing more general learning strategies as well (Hinze, Bunting & Pellegrino, 2009). The conclusion indicates that working memory was involved not only in reaching the final result, but also accompanied and supported the process of developing strategies to reach the outcome.

Good solvers can reach the deep meaning of a problem and detect characteristics and task similarities that are inaccessible to others. Efficient learners go beyond given information and discriminate between patterns stored in working memory (Lohman & Lakin, 2009). For them, encoding is rather a facilitator of the problem-solving than a part of the reasoning process, while poor solvers must invest great effort and have difficulties in finding the solution. A characteristic of efficient learners is their ability to process information selectively. Selective comparison means processing only a part of information in a particular situation, while selective combination is applied in tasks that require deductive reasoning, such as drawing a logical argument or a mathematical proof and involves strategically processing information in working memory. Individual differences in the ability to store information in working memory produce different outcomes. In a study about the effect of multiple representations on understanding and problem-solving abilities in physics, first-year high school participants who used more than one representation scored higher in performance than those who used a singular representation (Haratua & Judyanto, 2016). This points out that multiple representations could enhance students' comprehension, along with their problem-solving skills. Being able to build and store abstract models in working memory gives room to high-level functions and creates a powerful link between good knowledge and good reasoning (Markman & Gentner, 2001).

The fact that working memory plays an important role in developing efficient strategies was also investigated in the context of a deductive, nonverbal logical game in primary school children. The starting point of this research was a deductive version of the Mastermind game, used as a tool to explore the way people acquire complex logical skills and strategies (Verbrugge & Mol, 2008) that enable playing the game in the fastest possible way (Knuth, 1977; Kooi, 2005). The game was made suitable and implemented in a computerized adaptive learning system known as "Math Garden", applied in primary schools to train and monitor students' progress in logical reasoning (Van der Maas et al., 2010). A logical analysis of deductive Mastermind showed that the empirically established difficulty of items is related to the number of steps needed to find the solution (Gierasimczuk, Van der Maas & Raijmakers, 2013).

Empirical data of a large sample of school-aged children indicated, moreover, that working memory was specifically related to children's reasoning ability for this task.

In the current study, we are interested in testing the relevance of the Gierasimczuk et al. (2013) logical model for human reasoning. More in particular, we study the way strategies used by young adults evolve with experience, and the role working memory plays in this learning process.

Assessing reasoning strategies

There are different ways to assess strategies people use to solve a reasoning task. Below I will summarize some measures that have been used in cognitive research.

Verbal protocols consist of analyzing participants' verbal reports about the thoughts they generate to support their reasoning while solving a problem (Newel & Simon, 1972). The "thinking aloud" method provides information about the mental steps people take when trying to find a solution and helps identify the underlying cognitive processes that occur when solving a problem. Verbal protocols were also useful in exploring the degree of task difficulty and the extent to which people differed in their problem solving abilities, which may have had an impact on instruction (van Someren, Barnard & Sandberg, 1994). However, giving verbal protocols might interfere with solving the task and language may be difficult, especially for children.

Rule-assessment methodology (Siegler, 1976) explores the types of rules children used when having to judge a balance scale task, in which they were dealing with a combination of distance and weight information and had to predict how the balance will go to (left side down, right side down, stay balanced). Siegler constructed different problem types (with specific weight and distance configurations) to detect rules. The researcher discovered that children applied various more or less successful rules to solve balance scale problems, which led to different response patterns. The type of rules children used correlated with age, many young children relied on simple rules only using weight information, while the majority of older teenagers and adults used complex rules, integrating weight and distance information. The rule-assessment methodology provides an insight in children's reasoning process based on task performance only. This non-verbal character minimizes any kind of language bias. Finding similarities in the use of rules within a certain age group is an indicator of the consistency and stability of their behavior (Siegler & Chen, 2002).

Video-based eye-tracking method records eye movements in order to determine where the subject is looking at. Orienting eye gaze on a certain item in a problem-solving task shows where attention is focused and gives an indication of what people are thinking about (Rayner, 1998). There are two main features used to measure eye gaze: fixations and saccades. Fixations take place when eyes stay focused so that visual information can be stored, whereas saccades are rapid eye movements between fixations. The usefulness of the eye-tracking technique lies in the fact that the eye movements can reveal underlying reasoning strategies people apply when solving logical problems. In a study that investigated reasoning about mental states of the competitor in a strategic game (Meijering et al., 2012), researchers found that across trials, participants' looking pattern became more focused and converged towards a specific strategy that was predicted from theory. Authors discovered that in the beginning of the game, people tended to process relevant information through a perceptual straightforward forward progression of fixations, rather than a backward progression. When playing the game advanced, the likelihood of forward and backward reasoning was comparable. These findings show that people choose different strategies when performing a logical task, and these strategies had an adaptive function, changing over time in order to lead to a more efficient solution.

Another study (Evans, 1996) analyzed subjects' inspection times in a card game and found an association between the preferred cards and their corresponding inspection times, the selected cards being inspected for a longer time than the unchosen ones. In other words, attention selectively focused on items that were relevant to the task, while spending less time or ignoring those considered not important. Reasoning errors occurred either when people excluded relevant information or included irrelevant information in making a decision (Evans, 1996). People having the ability to look for central information and ignore or skip unimportant details performed better, which might have been due to their experience in similar tasks (Horn & Masunaga, 2006).

Eye-tracking is an efficient method that provides valuable information about the amount of time spent and the order in which relevant information is visually inspected for solving the task. In order to record eye movements, a task must be adapted to a computerized version with appropriate spatial layout.

Deductive Mastermind

In our design, we used the deductive version of the Mastermind game (DMM), known as the “Flower Code” in Math Garden (Van der Maas et al., 2010; Klinkenberg, Straatemeier & Van der Maas, 2011), an educational system for training of mathematics and logical reasoning. The game has many characteristics that make it appropriate for being played by children and is a relatively plain measure of logical reasoning: it has an easy to understand goal, it is nonverbal and focuses on abstract rules instead of background knowledge.

Figure 1. An item of the Deductive Mastermind (DMM) task, known as “Flower Code” in Math Garden. Figure is adapted from Gierasimczuk et al. (2012).

The DMM task in Math Garden consists of items (puzzles), with the following elements: a decoding board, a feedback instruction, the flowers to choose for the solution, the OK button, and a chronometer that counts 120 seconds-time measured by disappearing coins (Figure 1). In case of a correct answer, the number of left coins is added to the total score. However, if the response is not correct, the same number of coins is subtracted.

The goal is to find the correct sequence of flowers, based on the information given in the decoding board. Basically, each row of flowers is a conjecture that has a board on the right side. The board includes the feedback about the conjecture: one green dot means that one flower has the correct color and position for the solution, one orange dot means one flower has the correct color but it is in the wrong place, and one red dot means that one flower is not the solution at all.

As we are looking at 2-pin items (i.e., a code consisting of two flowers), there are two feedbacks for each line, in the following possible combinations: *orange-orange (oo)*, *red-red (rr)*, *green-red (gr)* and *orange-red (or)*.

The Analytic tableaux model

In the deductive version of the Mastermind game (Gierasimczuk, van der Maas, Raijmakers, 2013), for 2-pin items there were initially two more possibilities: green-green (*gg*) and green-orange (*go*). The green-green (*gg*) feedback was excluded because does not require any inference, thus it would be too easy. The green-orange (*go*) feedback contradicts itself, because green means that the flower should be in the solution and it is in the right position, while orange means that the flower is in the solution, but must be switched with the other flower, whose position we already know it is right. Therefore, both *gg* and *go* feedbacks were excluded from the game.

If we imagine the reasoning process as a tree with branches (mental steps), then each type of feedback requires branching to a different degree. For example, *oo* and *rr* feedbacks are conjunctions and do not require performing too many mental operations, *oo* giving the answer immediately, by just switching the flowers. In contrast, *gr* and *or* feedbacks are disjunctions and therefore require branching (Figure 2). Such an analysis is well explained by the Analytic tableaux model (Gierasimczuk, van der Maas & Raijmakers, 2013). The Analytic tableau method is a tree-shaped representation of the reasoning process that branches into smaller formulae until identifying a contradiction that makes impossible any further reduction. As an additional analysis step, the tree requires sometimes a flower that is not present in the conjectures at all.

Figure 2. Tableau model for 2-pin items. Each type of feedback has a different branching pattern (difficulty level). Adapted from Gierasimczuk et al. (2012).

Items difficulty

According to the Analytic tableau model, different types of feedback are of different difficulty level (Figure 2). Taking the number of required steps as an indicator of difficulty, the order of difficulty is as follows: *oo* (the easiest) < *rr* < *gr* < *or* (the most difficult). Easy 2-pin items are without *or* and *gr* feedback, or items without *or*, at least one *gr*, and all flowers of the solution must be found in at least one of the conjectures. Otherwise, items require more mental effort and are labeled as difficult. Item difficulty is determined by *or* steps, but only if processing *or* feedback is needed to solve the item, and excluding one color in the solution. For example, 2-pin items with only two colors available are considered easy (Gierasimczuk, van der Maas & Raijmakers, 2013). This classification in easy and difficult items is supported by empirical data.

Although the Analytic tableau method requires constructing the complete tree, processing all the information from all the conjectures is not always necessary. The length and complexity of the computation are dependent to some degree on the order in which the feedback is processed (Figure 3).

Figure 3. Two strategies that could be applied for one deductive Mastermind item. Adapted from Gierasimczuk et al. (2012).

For example (Figure 3), if starting with *gr* feedback (at the top of the item), we need to process all the conjectures to find the correct solution, as the *gr* feedback does not give the response immediately. However, if we begin with analyzing the *oo* feedback (at the bottom of the item), we have the solution straight away (in this case, *tulip-sunflower*), and processing *gr* feedback becomes irrelevant. In the current research, we will test predictions derived from the Analytic tableau model.

Current study

We aim to explore in an eye-tracking study how people reason while performing the deductive reasoning task and whether there is an improvement in the use of strategies with increasing experience. To be more specific, we tested predictions derived from the study of Gierasimczuk, Van der Maas & Raijmakers (2013).

First of all, we tested the prediction of the model that items had a different difficulty level depending on type of feedback necessary to process. This is an important model prediction, because the difficulty level indicates the complexity of mental steps people follow in their reasoning: the more difficult the feedback, the more sophisticated the strategies must be. Moreover, it is predicted that students improve their efficiency over time. Secondly, we wanted to explore whether the difficulty depends to some extent on the order in which the conjectures were processed. To this end, we introduced items that only differed in the order in which information was presented. Finally, it is predicted that working memory plays a role in developing more or less complex strategies.

Unlike the study of Gierasimczuk, van der Maas & Raijmakers (2013), we designed our study involving first-year students as participants instead of 5-12 year children, in order to explore the way older participants reason and apply strategies over time.

Research question 1 explores whether participants become more efficient in solving the items over blocks. We expect that improvement becomes evident by a reduced number of fixations, a decreased duration and a better accuracy in performing the task.

Research question 2 explores whether items have different difficulty levels, as measured by time spent on processing items and accuracy, dependent on the type of feedback information involved. We expect that processing easy items requires shorter time and provides a higher accuracy. For difficult items, we expect vice versa.

Research question 3 explores the role of working memory capacity in the type of strategies young adults apply after experience. We expect that a better working memory capacity predicts the use of more efficient strategies. Good memorizers possess a higher reasoning ability, tend to use more complex strategies and are able to find shortcuts in solving problems over time, which makes them faster and more efficient. More specifically, we expect that participants with a good working memory spend less time, make less fixations and are more accurate than poorer memorizers in performing the task after experience.

Research question 4 explores whether the strategy participants learn by experience is to become more selective in evaluating information (i.e. focusing on the easiest, necessary feedback to process information), leading to a faster solution. We expect that with experience, the order in which information is processed gets more dependent on the feedback information and less dependent on the order in which information is presented. More specific predictions are provided in the methods and results sections.

Methods

Participants

The sample consisted initially of 26 participants, but 3 were excluded due to missing data in the eye-tracking measurements. The remaining 23 participants had an age range between 17.67 and 27.97 ($M = 20.87$, $SD = 2.41$). The distribution by gender was divided in 78.3% female ($n = 18$) and 21.7% male ($n = 5$). Among the total number of participants, 56.5% ($n = 13$) were Pedagogy students, while 43.5% ($n = 10$) were Psychology students. With respect to the country of origin criteria, 69.6% ($n = 16$) were from The Netherlands, and 30.4 % ($n = 7$) came from other countries. Students took part in the experiment to receive course credits. The project was approved by the ethics committee of the Department of Child and Education, University of Leiden. All participants gave active informed consent preceding the experiment.

Materials

Item types.

The experiment was structured in a learning block and two experimental blocks. For the learning phase, we constructed 13 items (L-items, Appendix 1) and its goal was to make sure participants understood the task. Subsequently, we designed 16 experimental items for the first block (E1) and 16 experimental items for the second block (E2). The items used in the experimental blocks were 2-pin items with 5 rows, constructed as an adapted version of the task

that was suitable for eye-tracking. Items were equivalent regarding the informative value of the conjectures and the type of feedback. The only discriminant element was the type of flowers; novel flowers were used in the second block. We will discuss the experimental items, which were 4 items of 4 different types.

Items 1-4 (T1-type) were composed of the same lines, with the same feedback (*rr*, *gr* and *or*), but conjectures were displayed in a different order with different flower combinations (*gr-or-gr-gr-rr*; *gr-or-rr-gr-gr*; *gr-gr-rr-gr-or*; *rr-gr-gr-gr-or*), and all the flowers were in at least one of the conjectures. If processed starting with *gr-rr* or *gr-gr* feedback, the items were considered easy. In contrast, if the *or* feedback is processed firstly, items were considered difficult (Appendix 2).

Items 5-8 items (T2-type) were composed of the same lines, with the same feedback (*oo*, *gr*, *or*), but conjectures were presented in a different order with different flower combinations (*gr-or-gr-gr-oo*; *gr-or-oo-gr-gr*; *gr-gr-oo-gr-or*; *oo-gr-gr-gr-or*), and all the flowers were in at least one of the conjectures. Items T2-type were considered easy selective, if processed starting with the *oo* feedback, as that provided the solution immediately by just switching the flowers. However, if starting with analyzing the *or* feedback, the item was considered difficult (Appendix 3).

Items 9-12 (T3-type) were composed of the same lines, with the same feedback (*rr-rr-gr*) but conjectures were displayed in a different order with different flower combinations (*rr-rr-gr-gr-gr*; *gr-gr-gr-rr-rr*). Items T3-type contained a flower that was not included in none of the conjectures and were considered difficult, no matter the direction of processing (Appendix 4).

Items 13-16 (T4-type) were constructed of the same lines, with the same feedback (*rr-gr-or*) but conjectures were placed in different order with different flower combinations (*gr-gr-or-gr-rr*; *gr-gr-rr-or-gr*; *rr-gr-gr-or-gr*; *or-gr-rr-gr-gr*). Item 4-type had an external flower that was not included in none of the conjectures. If analyzed starting with the *gr-gr* or *gr-rr* feedback, the items were considered easy. Otherwise, if the *or* feedback was the first one to be processed, the item was considered difficult (Appendix 5).

Duration represented the total amount of time spent on reading different conjectures of an item, expressed in milliseconds. A lower duration indicated a shorter time allocated to inspect a specific conjecture, whereas a higher duration showed more time spent on that conjecture.

However, participants could not spend more than two minutes in total on an item. Continuously decreasing time was displayed during presentation.

Eye movements were quantified as the total number of fixations within the areas of interest (conjectures) of an item. A lower amount of fixations indicated that participants spent less time focusing on that area, while a higher number of fixations indicated an increased focus.

Accuracy was expressed in sum scores from 0 to 16 for each block. A sum score was calculated by adding up all the correct answers, 0 indicated lower performance, while 16 showed optimal performance.

Visual working memory was measured by a version of the Chessboard working memory task (Dovis et al., 2012), which assesses the capacity to both store and manipulate relevant visuospatial information. Participants were presented a chessboard that consisted of green and blue squares. The computer highlighted and acoustically signalized certain green and blue squares, in a random order. The task was to remember and then to reproduce the correct order in which the computer displayed the squares, but starting with the green boxes. Participants had to mouse-click first the green squares in the correct order, and then the blue squares, in the correct order as well. The score was calculated by the mean sequence length of the last 10 trials. A lower score showed a poorer WM capacity, and a higher score indicated a better WM performance (Dovis et al., 2012).

Apparatus.

Eye-tracker. Eye-tracking is an unobtrusive technique that has been consistently used not only in reading tasks, but also in investigating logical reasoning. Eye movements were recorded by an EyeLink-1000 eye tracker (SR Research Ltd., Ontario, Canada), at a sample rate of 500 Hz. Students were seated on a comfortable chair approximately 55 cm from a 17-in. computer monitor, which subtended an approximate 27° x 34° visual angle. A chin support device attached to the desk enabled participants to maintain their head in a stable position. Before the data collection, a five-point calibration technique was used to calibrate each participant's point of gaze. If necessary, the calibration procedure was repeated until the recorded point of gaze reached the best possible quality. The experimental session only started after the calibration was successful. After that, the total 45 items (13 L, 16 E1 and 16 E2) were displayed one at a time for a maximum of 2 minutes each.

Procedure

Participants were recruited through the University of Leiden in two ways: by posting flyers in visible places at the Faculty of Pedagogy and Psychology, and via the Sona system (through which students get course credits for participating in studies). In the testing day, before starting the experiment, they read the information letter about the tasks they were supposed to perform and signed the informed consent. Subsequently, participants performed the deductive reasoning task (with a 2-minute break between blocks L, E1, E2, when they could move their head), while their eye movements were recorded by the eye-tracker. Calibration took place before each block.

The differences in duration, the number of fixations recorded by the eye-tracker and the accuracy of items were recorded. After finishing the deductive reasoning task, participants completed the working memory task on the computer. In the end, they were debriefed and asked whether they had further questions or comments about the experiment. In case they did, the researcher answered them as accurately as possible.

Data analysis

We structured our analysis on two main aspects. The first set of research questions focused on general changes over blocks and differences between item types (T1, T2, T3, T4). The second set of research questions focused on differences in the looking pattern between a specific strategy versus others within different item types (*or* in T1, *oo* in T2, *rr* in T3, *or* in T4) over blocks. For testing the hypotheses derived from the first three research questions (1.1-3.3), a repeated-measures MANCOVA was performed, with the number of fixations, duration and accuracy as dependent variables, and item types (with four levels) and block (with two levels) as categorical predictors. Working memory capacity was taken into account as a covariate.

Research question 1 explores whether participants become more efficient in solving the items over blocks, and this improvement is measured by a reduced number of fixations, a decreased duration and a better accuracy in performing the task. We expected thus a main effect of block on the number of fixations, duration and accuracy.

Hypothesis 1.1: Number of fixations decreases over blocks.

Hypothesis 1.2: Time spent on looking at relevant information item decreases over blocks.

Hypothesis 1.3: Responses become more accurate over blocks.

Research question 2. Items have different difficulty levels, determined by the type of feedback. The difficulty level is measured by the time spent on processing items and accuracy. A main effect of item type on fixations, duration and accuracy was expected.

Hypothesis 2.1: Number of fixations is larger in difficult items.

Hypothesis 2.2: Solvers spend more time looking at relevant information in difficult items.

Hypothesis 2.3: Accuracy is higher in easier items.

Research question 3 investigates whether a better working memory capacity predicts the use of more efficient strategies over blocks. We expected that participants with a good WM spent less time, made less fixations and were more accurate than poor memorizers in the second block. Therefore, it was anticipated an interaction between working memory and learning.

Hypothesis 3.1: In the second block, good memorizers make fewer fixations than participants with a lower WM capacity.

Hypothesis 3.2: In the second block, good solvers with a good WM capacity spend less time on inspecting the items.

Hypothesis 3.3: In the second block, good memorizers have higher accuracy than poorer memorizers.

Research question 4 investigates whether the learning process occurs by becoming more selective in evaluating information (i.e. focusing on the easiest, necessary feedback to process information). We anticipated that participants apply two ways of processing information: based on order (top-bottom or bottom-up), and based on selecting a specific feedback (Table 1). The last way we expected to be more efficient and to play a more important role in the learning process. Thus, looking times are dependent on the specific order people process information.

Hypothesis 4.1: For T1 items that contain *or*, *gr* and *rr* feedback, participants look relatively less at or feedback over blocks.

Hypothesis 4.2: For T2 items that contain *oo*, *gr* and *or* feedback, participants look relatively longer at oo feedback over blocks.

Hypothesis 4.3: For T3 items that contain *rr* and *gr* feedback, participants look relatively longer at rr feedback over blocks.

Hypothesis 4.4: For T4 items that contain *or*, *rr* and *gr* feedback, participants look relatively less at or feedback over blocks.

In order to test the hypotheses 4.1-4.5, a second series of models was tested, which focused on the types of strategies students used to solve the task: based on the order in which feedback was presented versus selecting a specific feedback within different item types (*or* in T1, *oo* in T2, *rr* in T3, *or* in T4) over blocks. For each item type (T1, T2, T3, T4) separately, a repeated-measures MANCOVA was conducted, with relative duration as dependent variable, block (E1/E2) and individual items (four levels, a-d) as within factors, the type of feedback (*or* versus other feedback in T1, *oo* versus other feedback in T2, *rr* versus other type of feedback in T3, and *or* versus other type of feedback in T4) as categorical predictors, and working memory as a covariate.

Table 3 summarizes the types of items and types of strategies that could be used by participants according to the predictions of the model.

Table 1

Types of strategies and item types

Item types	Items description	Strategy	
		Processing all information	Selective processing
Type 1	<i>or</i> feedback without external flower	difficult	easy
Type 2	<i>oo</i> feedback without external flower	difficult	easy
Type 3	<i>rr</i> feedback with external flower	difficult	difficult
Type 4	<i>or</i> feedback with external flower	difficult	easy

Results

Research question 1. The first research question explored whether participants became more efficient in solving the items over blocks, and this improvement became evident by a reduced number of fixations, a decreased duration and a better accuracy in performing the task.

A repeated-measures MANCOVA, with number of fixations, duration and accuracy as three dependent variables, and block (E1/E2), item type (T1, T2, T3, T4) as categorical predictors and working memory as a covariate was calculated to answer hypotheses 1.1-3.3. We tested all main and two-way interactions effects of block, item type and working memory

capacity on the number of fixations, duration and accuracy. We expected thus a main effect of block on the number of fixations, duration and accuracy.

Descriptive statistics for the dependent variables are shown in Table 2. Table 3 summarizes all the needed information about the MANCOVA analysis.

Table 2

Descriptive statistics for all key-variables per groups and subgroups

Variable	<i>N</i>	<i>Block E1</i>		<i>Block E2</i>	
		<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>
Duration (ms)	23	10902	5240	7900	4157
Number of fixations	23	48.58	20.18	35.37	14.28
Accuracy	23	.79	.20	.92	.12
% Duration <i>or</i> feedback T1	23	.17	.06	.16	.06
% Duration <i>oo</i> feedback T2	23	.43	.18	.52	.22
% Duration <i>rr</i> feedback in T3	23	.41	.14	.49	.21
% Duration in <i>or</i> feedback in T4	23	.29	.05	.29	.05
Working memory	23	6.49	1.09		

Table 3

Repeated-measures MANCOVA of duration, fixations and performance. Only significant results are displayed

	<i>SS</i>	<i>df</i>	<i>MS</i>	<i>F</i>	<i>p</i>	η^2
<i>Duration (ms)</i>						
Block	412318143	1	412318143	41.00	.00	.66

Item type	525203246	1.92	274089875	7.82	.00	.32
Block x WM	56101198	1	56101198	5.64	.03	.21
<i>Number of fixations</i>						
Block	7984.13	1	7984.13	41.00	.00	.66
Item type	11742.82	1.98	5930.26	9.79	.00	.32
Block x WM	1098.73	1	1098.73	5.64	.03	.21
<i>Accuracy</i>						
Block	.72	1	.72	16.86	.00	.45
Item type	2.65	1.79	1.48	21.17	.00	.50
Block x Item type	.51	3	.17	9.13	.00	.30

Hypothesis 1.1. Number of fixations decreases over blocks. Findings show a significant main effect of block, $F(1, 21) = 41.00$, $p < .001$, $\eta^2 = .66$ (Block E1: $M = 48.58$, $SD = 20.18$; Block E2: $M = 35.37$, $SD = 14.28$). Participants fixated the items less in Block E2, which confirms hypothesis 1.1 (Figures 4 and 5).

Figure 4. Effect of block on the number of fixations. Number of fixations decreased over blocks.

Figure 5. Number of fixations as recorded by the eye-tracker for a typical participant. This figure illustrates a more focused eye-gaze, thus a more efficient strategy in Block E2 (191 fixations at the conjectures in Block E1 versus 23 fixations in Block E2).

Hypothesis 1.2. Time spent on looking at relevant information decreases over blocks. Findings show a significant main effect of block, $F(1, 21) = 41.54, p < .001, \eta^2 = .66$ (Block E1: $M = 10902, SD = 5240$; Block E2: $M = 7900, SD = 4157$), solvers processed the items shorter in Block E2 (Figure 6), which confirms hypothesis 1.2.

Figure 6. Effect of block on duration. Time spent on processing relevant information of the item decreased over blocks.

Hypothesis 1.3. Responses become more accurate over blocks. A significant main effect of block is present, $F(1, 21) = 16.86, p < .001, \eta^2 = .45$ (Block E1: $M = .79, SD = .20$; Block E2: $M = .92, SD = .12$). Accuracy increased in Block E2, which confirms hypothesis 1.3 (Figure 7).

Figure 7. Effect of block on accuracy. Accuracy improved in Block E2.

Research question 2 explored whether items had different difficulty levels, as measured by time spent on processing items and accuracy, dependent on the type of feedback information involved.

Hypothesis 2.1. Number of fixations is larger in difficult items. There is a significant main effect of item type, solvers focused to a different degree on various types of items, $F(1.98, 41.58) = 9.79, p < .001, \eta^2 = .32$. Pairwise comparisons (Bonferroni corrected) indicate significant differences between T1 ($M = 46.48, SD = 21.92$) and T2 items ($M = 30.80, SD = 23.20$), $t(22) = 5.46, p < .001$, between T2 ($M = 30.80, SD = 23.20$) and T4 ($M = 51.97, SD = 18.17$), $t(22) = -3.84, p < .001$, between T3 ($M = 38.63, SD = 18.56$) and T4 ($M = 51.97, SD = 18.17$), $t(22) = -3.59, p < .001$, as well as between T1 ($M = 46.48, SD = 21.92$) and T3 ($M = 38.63, SD = 18.56$), $t(22) = 2.23, p = .04$. The highest amount of fixations was on T4 and T1 items (both containing *or* feedback), and the lowest amount of time was allocated to T2 items (*rr* feedback) and T3 (*oo* feedback), which confirms hypothesis 2.1.

Hypothesis 2.2. Solvers spend more time looking at relevant information of the item in difficult items. Findings show a significant main effect of item type, $F(1.92, 40.24) = 7.82, p < .001, \eta^2 = .27$. Pairwise comparisons (Bonferroni corrected) reveal a similar pattern as on fixations; significant differences are present between T1 ($M = 10531, SD = 6237$) and T2 items

($M = 7121$, $SD = 6292$), $t(22) = 5.02$, $p < .001$, between T2 ($M = 7121$, $SD = 6292$) and T4 ($M = 11446$, $SD = 4183$), $t(22) = -3.35$, $p < .001$, between T3 ($M = 8506$, $SD = 4538$) and T4 ($M = 11446$, $SD = 4183$), $t(22) = -3.56$, $p < .001$, as well as between T1 ($M = 10531$, $SD = 6237$) and T3 ($M = 8506$, $SD = 4538$), $t(22) = 2.36$, $p = .03$. The processing time was longer in T4 and T1 items (that contain *or* feedback), and shorter in T2 and T3, with *oo* and respectively *rr* feedback, which confirms hypothesis 2.2.

Hypothesis 2.3: Accuracy was higher in easier items. There is a significant main effect of item type in accuracy, $F(1.79, 37.58) = 21.18$, $p < .001$, $\eta^2 = .50$, accuracy was different in various items, which confirms hypothesis 2.3. Pairwise comparisons (Bonferroni corrected) indicate significant differences between items T1 ($M = .90$, $SD = .20$) and T4 ($M = .65$, $SD = .25$), $t(22) = 4.80$, $p < .001$, between items T2 ($M = .91$, $SD = .19$) and T4 ($M = .65$, $SD = .25$), $t(22) = 4.82$, $p < .001$, as well as between T3 ($M = .96$, $SD = .07$) and T4 ($M = .65$, $SD = .25$), $t(22) = 6.09$, $p < .001$, with a higher accuracy in T3, T2 and T1 items, and a lower accuracy in T4 items. There is also an interaction effect between block and item type, $F(1.89, 39.66) = 9.13$, $p < .001$, $\eta^2 = .30$, the greatest improvement in accuracy with experience being noticed in T4 items.

Research question 3. A better working memory capacity predicts the development of more efficient strategies. A repeated-measures MANCOVA was performed, with the number of fixations, duration and accuracy as dependent variables, block as predictor and WM as a covariate. All main effects and two-way interactions were tested.

Hypothesis 3.1: In the second block, good memorizers decrease the number of fixations more than lower memorizers (interaction effect of WM and block on fixations). There is an interaction effect of block and WM, $F(1, 21) = 5.64$, $p = .03$, $\eta^2 = .21$. Good memorizers focused their eyes less over blocks, which confirms hypothesis 3.1 (Figure 8).

Figure 8. Effect of WM on the difference in fixations between blocks E1 and E2.

Hypothesis 3.2. In the second block, good solvers with a good WM capacity decrease inspection time of relevant information more than in the first block. There is an interaction effect between block and working memory, $F(1, 21) = 5.65$, $p = .03$, $\eta^2 = .21$. Higher memorizers spent less time on processing relevant information over blocks, which confirms hypothesis 3.2. (Figure 9).

Figure 9. Effect of WM on the difference in duration between blocks E1 and E2.

Hypothesis 3.3. In the second block, good memorizers increase accuracy more than those with a poorer WM capacity. However, no interaction is present between WM and learning, thus hypothesis 3.3 is not confirmed.

Research question 4. Learning process occurs by becoming more selective in evaluating information (selecting the easiest feedback). To test this research question, we used the second set of analyses, focused on the type of feedback within item types. For each item type (T1, T2, T3, T4) separately, a repeated-measures MANCOVA was conducted, with relative duration as dependent variable, block (E1/E2) and individual items (four levels, a-d) as within factors, the type of feedback (*or* versus other feedback in T1, *oo* versus other feedback in T2, *rr* versus other type of feedback in T3, and *or* versus other type of feedback in T4) as categorical predictors, and working memory as a covariate.

Hypothesis 4.1: For T1 items, which contain *or*, *gr* and *rr* feedback, people look relatively less at *or* feedback over blocks. That is, they choose an easier and more efficient strategy with experience, by spending shorter time or even ignoring *or* feedback, which is considered difficult, according to the Analytic tableau model. There is no main effect of block, students spending a comparable amount of time looking at *or* feedback over blocks, thus hypothesis 4.1 is not confirmed. A main effect of item is present, $F(3, 66) = 19.14, p < .001, \eta^2 = .47$. Solvers looked for a different amount of time at *or* feedback when this conjecture was placed in a different order. Pairwise comparisons (Bonferroni corrected) indicate significant differences between items T1a ($M = .20, SD = .09$) and T1c ($M = .10, SD = .08$), $t(22) = 4.83, p < .001$, between T1a ($M = .20, SD = .09$) and T1d ($M = .11, SD = .08$), $t(22) = 3.87, p < .001$, between T1b ($M = .26, SD = .11$) and T1c ($M = .10, SD = .08$), $t(22) = 5.77, p < .001$, between T1b ($M = .26, SD = .11$) and T1d ($M = .11, SD = .08$), $t(22) = 5.57, p < .001$, as well as between T1a ($M = .20, SD = .09$) and T1b ($M = .26, SD = .11$), $t(22) = -2.20, p = .04$. These differences indicate that a longer time was spent on *or* feedback when it was placed at the top of the item-see T1b items (Figure 10).

Figure 10. Effect of block (E1/E2) and item (a-d) on T1 items. Participants spent a comparable amount of time on *or* feedback in both blocks.

Hypothesis 4.2: For T2 items, which contain *oo*, *gr* and *or* feedback, participants look relatively longer at *oo* feedback over blocks. There is a significant main effect of block, $F(1, 21) = 7.60$, $p = .01$, $\eta^2 = .27$ (Block E1: $M = .43$, $SD = .18$; Block E2: $M = .52$, $SD = .22$). Participants spent significantly more time looking at *oo* feedback in Block E2, which confirms hypothesis 4.2. There is a main effect of item as well, $F(1.83, 38.41) = 10.32$, $p < .001$, $\eta^2 = .33$, participants spending a different amount of time when *oo* feedback was placed in a different order. Pairwise comparisons (Bonferroni corrected) show significant differences between items T2a ($M = .36$, $SD = .28$) and T2d ($M = .66$, $SD = .23$), $t(22) = -4.14$, $p < .001$, between T2b ($M = .43$, $SD = .22$) and T2d ($M = .66$, $SD = .23$), $t(22) = -3.58$, $p < .001$, as well as a marginally significant difference between items T2a ($M = .36$, $SD = .28$) and T2c ($M = .46$, $SD = .27$), $t(22) = -2.08$, $p = .05$. Participants spent more time on *oo* feedback when this conjecture was placed at the top of the item-see T2d items (Figure 11).

Figure 11. Effect of block (E1/E2) and item (a-d) on T2 items. Percentage duration of looking at *oo* feedback increased in Block E2.

Hypothesis 4.3: For T3 items, which contain *rr* and *gr* feedback, participants look relatively longer at *rr* feedback over blocks. There is a significant main effect of block, $F(1, 21) = 7.13, p = .01, \eta^2 = .25$ (Block E1: $M = .41, SD = .14$; Block E2: $M = .49, SD = .21$), solvers inspected *rr* feedback relatively longer in Block E2, which confirms hypothesis 4.3. In addition, a main effect of item is present, $F(3, 63) = 15.59, p < .001, \eta^2 = .43$. Pairwise comparisons (Bonferroni corrected) indicate significant differences between items T3a ($M = .52, SD = .20$) and T3c ($M = .32, SD = .19$), $t(22) = 6.13, p < .001$, as well as between T3b ($M = .55, SD = .18$) and T3c ($M = .32, SD = .19$), $t(22) = 5.80, p < .001$. Participants looked longer at *rr* feedback when this conjecture was placed at the top of the item-see T3b and T3a items (Figure 12).

Figure 12. Effect of block (E1/E2) and item (a-d) on T3 items. Percentage duration of looking at *rr* feedback increased over blocks.

Hypothesis 4.4: For T4 items, which contain *or*, *rr* and *gr* feedback, there is neither main effect of block, nor main effect of item, thus hypothesis 4.4 is not confirmed.

Discussion

The present study had four important aims. The first aim was to explore whether students improve their efficiency in solving a deductive reasoning task with experience. The second aim was to test a prediction of the Gierasimczuk, Van der Maas & Raijmakers (2013) model, which anticipates different levels of difficulty in different item types, dependent on the feedback information involved. The predicted difficulty of conjectures in the items is as follows: the easiest to process are conjectures with feedback *oo*, which are followed by *rr*, *gr*, and conjectures with feedback *or*, which are considered the most difficult to process. The third aim was to analyze whether a better working memory capacity predicts the application of more efficient strategies. The fourth aim was to explore whether the learning process occurs by becoming more selective in evaluating relevant information present in the item.

With respect to the first aim, participants improved their efficiency with experience, and this could be reflected in the way they handled the task and in their final performance. Students became significantly faster and more accurate between blocks, suggesting that an underlying learning process occurred over time. This conclusion is consistent with other findings, which stated that people simplified their reasoning over time, in order to become more efficient (Hinze, Bunting & Pellegrino, 2009).

Regarding the second aim, we found that items differed in the difficulty level, with more time spent on difficult items that contain *or* feedback and better accuracy in easy items (*oo* and *rr* feedback). The findings are consistent with the Analytic tableaux model and confirm the predictions. Discovering that participants perceived items as having different difficulty levels is relevant as it gives an insight in the cognitive processes that take place while reasoning. According to the Analytic tableau model (Gierasimczuk et al., 2013), item difficulty indicates the complexity of mental steps complexity that reasoners have to follow: the more challenging the item is, the more sophisticated the reasoning strategy will become.

With respect to the third aim, we found that participants with a good working memory capacity needed shorter time to process item information than poorer memorizers in the second block relative to the first block. The conclusion is consistent with previous findings according to which a good working memory enabled solvers to search for mental shortcuts in order to arrive rapidly to a solution. By simplifying their reasoning, people reduced the cognitive load and provided more efficient responses with experience (Hinze, Buntin & Pellegrino, 2009). However, even though in our study good memorizers became faster with learning, they did not score higher in accuracy than poorer memorizers across trials. This indicates that a higher WM capacity did not guarantee a better performance per se, but rather choosing a faster strategy, which was not necessarily the most effective. The result points out that working memory played a more important role in developing a reasoning strategy than in performance.

Concerning the fourth aim, participants looked for a relatively higher amount of time at *oo* and *rr* feedback across trials, suggesting the preference for a specific strategy. Choosing to focus more on *oo* and *rr* feedback in the second block indicates that selective information processing increased with learning. With experience, participants did not spend unnecessary time focusing on difficult feedback anymore and concentrated on the most relevant information contained in the item, which ensured a straightforward solution (i. e. in *oo* feedback). We also expected that participants would choose to spend less time on *or* feedback in the second block, as it is considered the most difficult in the Gierasimczuk et al. (2013) model. However, analysis indicates that the focus on *or* feedback was quite similar in both blocks, even though we already found that *or* feedback in T1 and T4 items needed the longest time to process. The fact that solvers kept looking at *or* feedback instead of choosing an easier way (i. e., looking at *gr-gr* feedback) might be explained by the absence of *oo* feedback in T1 and T4 items. As a result,

students could not figure out another simpler way to solve the items. Besides developing a selective strategy, students also had a tendency to process from top to bottom in performing the task. Eye-tracking records revealed that they looked longer at the relevant feedback (specifically *oo* and *rr*) when it was placed at the top of the item than when this conjecture was placed at the bottom of the item. This indicates that order of information played a role, by focusing more on a specific strategy; in this case, looking from top to bottom. This means a forward direction of reasoning, namely processing information from the first premise to the last one. The conclusion is in line with other findings that showed that forward reasoning represented a more familiar and prevalent strategy in first-year university students (Meijering et al., 2012).

Another important aspect was the absence of a significant interaction between the type of feedback (*or*, *oo*, *rr*) and working memory. Focusing for a longer or shorter time on a specific type of feedback did not depend on WM levels. Even if good memorizers developed better strategies, there is no evidence to conclude that choosing a more efficient strategy is due to the fact that high WM people were better at selecting relevant information.

Implications.

The current study may provide a useful insight in the way young adults learn to develop and adapt their strategies with experience in a deductive reasoning task. The conclusions derived from the study have both theoretical and practical consequences. The first important finding is that people became more efficient with experience. Identifying a main effect of block leads to the conclusion that deductive reasoning can be trained in just a few items. This has at least two implications for educational practice. Firstly, that if the result would also hold for children, then using deductive games as an instructional technique might be likely successful in the school curriculum. A way to stimulate deductive reasoning skills might be to play stimulating computerized games similar to those in Rekentuin (Van der Maas, H., Klinkenberg, S., Straatemeier, M., 2010), that are adapted to children's age and own individual level. Secondly, in the classroom, teachers can expose pupils from an earlier age to solving deductive items. There is evidence indicating that children are able to reason and solve problems since infancy (Woodward, 2009), and during the preschool years, their logical thinking and reasoning abilities develop consistently (Amsterlaw & Wellman, 2006). These findings support the need to start training the deductive reasoning skills as early as possible.

The second important finding confirms the prediction of Gierasimczuk et al. (2013) regarding the level of item difficulty. This may have implications in constructing deductive reasoning games varying in the level of difficulty for children with different educational needs. A better understanding of the manner children handle tasks of different difficulty level could lead to constructing new items more carefully.

The third key-finding is that students developed strategies across trials, becoming more selective in processing relevant information. Being more focused and efficient in what is important to process in a task is a crucial skill to learn, because it facilitates success later in life (Siegler & Svetina, 2006). Not only in the academic setting, but also in real-life, the better people can discriminate between important versus non-important information, the more adapted they would become to every situation they encounter.

The fourth finding is that a good working memory facilitated more the improvement of strategies than a poorer WM, with learning. Good memorizers applied more efficient strategies towards the end of the task and learned more than their peers across trials. The relationship that exists between WM and types of strategies claims for the necessity to design training programs for individuals with a lower WM, in order to improve their capacity to construct more efficient reasoning strategies. Empirical evidence indicated that WM training can be used as an instrument to increase higher-level cognitive functions across a wide range of populations (Morrison & Chein, 2011).

Limitations.

One limitation of the present study may be the sample size ($N = 23$), which did not ensure drawing firm conclusions about the findings. As a matter of fact, our study might be suffering from a low power, drawback that could be overcome by increasing the number of participants. Another disadvantage may be the structure of the sample, which consisted of mainly first-year students and female participants. This fact did not allow a too high variability in data for a more accurate analysis. On the other hand, our experiment is different from the study of Gierasimczuk, van der Maas & Raijmakers (2013) with respect to the sample structure. In Rekenhuis, researchers used primary school Dutch children as participants, thus it is expected that they received the same level of education. However, our participants differed greatly with respect to their cultural and educational background. Students came from different countries (The

Netherlands, Greece, Germany, Hungary, Indonesia, China), which might imply a higher variability in the way they approached the task at hand.

Further research.

Further studies may focus more firstly on constructing different tasks for exploring the deductive reasoning process. Secondly, a comparative design between children and young adults could investigate the way different age groups handle with the same type of task. In doing this, a more homogenous structure in the adult sample with respect to their educational background would be more appropriate for a developmental study. Lastly, it would be interesting to design a deductive reasoning training of an increased length. This way, one could test whether participants are adapting fully to an information selection strategy.

References

- Amsterlaw, J., Wellman, H. M. (2006). Theories of mind in transition: A microgenetic study of the development of false belief understanding. *Journal of Cognition and Development*, 7(2), 139-172.
- Bhat, M. A. (2016). The predictive power of reasoning ability on academic achievement. *International Journal of Learning, Teaching and Educational Research*, 15(1), 79-88.
- Copeland, D. E., & Radvansky, G. A. (2004). Working memory and syllogistic reasoning. *The Quarterly Journal of Experimental Psychology*, 57(8), 1437-1457.
- Cowan, N. (1995). Attention and memory: An integrated framework. Oxford Psychology Series (No. 26). New York: Oxford University Press.
- Dovis, S., Van der Oord, S., Wiers, R.W., Prins, P. J. M. (2012). Can motivation normalize working memory and task persistence in children with attention-deficit/hyperactivity disorder? The effects of money and computer-gaming. *Journal of Abnormal Child Psychology*, 40, 669–681.
- Evans, J. S. B. T. (1996). Deciding before you think: Relevance and reasoning in the selection task, *British Journal of Psychology*, 87(2), 223-240.
- Gierasimczuk, N., van der Maas, H. L. J., & Raijmakers, M. E. J. (2013). An analytic tableaux model for deductive Mastermind tested with a massively used online learning system. *Journal of Logic, Language and Information*, 22(3), 297-314.
- Gierasimczuk, N., Van der Maas, H. L. J., & Raijmakers, M. E. J. (2012). Logical and psychological analysis of deductive Mastermind. *Journal of Logic, Language and Information*, 883, 1-13.

- Gilhooly, K. J. (2004). Working memory and reasoning. In J. P. Leighton & R. J. Stemberg (Eds.), *The nature of reasoning* (pp. 49-77). New York, NY: Cambridge University Press.
- Gilhooly, K. J., Phillips, L. H., Whynn, V., Logie, R. H., & Della Sala, S. (1999). Planning processes and age in the five-disc Tower of London task. *Thinking and reasoning*, 5(4), 339-361.
- Haratua, T. M. S., Judyanto, S. (2016). Representations based physics instruction to enhance students' problem solving. *American Journal of Educational Research*, 4(1), 1-4.
- Hinze, S. R., Bunting, M. F., & Pellegrino, J. W. (2009). Strategy selection for cognitive skill acquisition depends on task demands and working memory capacity. *Learning and Individual Differences*, 19, 590-595.
- Holyoack, K. J., & Morrison, R. G. (2005). Thinking and reasoning: A reader's guide. In K. Holyoack & B. Morrison (eds.), *The Cambridge Handbook of Thinking and Reasoning*. Cambridge University Press.
- Horn, J., & Masunaga, H. (2006). A merging theory of expertise and intelligence. In K. A. Ericsson, N. Charness, P. J. Feltovich, & R. R. Hoffman (Eds.), *The Cambridge handbook of expertise and expert performance* (pp. 587-612). New York, NY: Cambridge University Press.
- Johnson- Laird, P. N. (2004). Mental models and reasoning. In J. P. Leighton & R. J. Stemberg (Eds.), *The nature of reasoning* (pp. 49-77). New York, NY: Cambridge University Press.
- Klinkenberg, S., Straatemeier, M., Van der Maas, H. L. J. (2011). Computer adaptive practice of Maths ability using a new item response model for on the fly ability and difficulty estimation. *Computers & Education*, 57, 1813-1824.
- Knuth, D. (1977). The computer as Master Mind. *Journal of Recreational Mathematics*, 9, 1-6.

- Kooi, B. (2005). Yet another mastermind strategy. *ICGA Journal*, 28(1), 13-20.
- Lohman, D. F., & Lakin, J. M. (2009). Reasoning and intelligence, in R. J. Sternberg & S. B. Kaufman (Eds.), *Handbook of Intelligence* (2nd ed.). New York, NY: Cambridge University Press.
- Lustig, C., May, C. P., & Hasher, L. (2001). Working memory span and the role of proactive interference. *Journal of Experimental Psychology*, 130 (2), 199-207.
- Markman, A. B., & Gentner, D. (2001). Thinking. *Annual Review of Psychology*, 52, 223-247.
- Meijering, B., Van Rijn, H., Taatgen, N. A., Verbrugge, R. (2012). What eye movements can tell about Theory of mind in a strategic game. *PloS one*, 7(9): doi:10.1371/journal.pone.0045961
- Morisson, A. B., Chein, J. M. (2011). Does working memory training work? The promise and challenges of enhancing cognition by training working memory. *Psychonomic Bulletin & Review*, 18 (1), 46-60.
- Newell, A., & Simon, H. A. (1972). *Human problem solving*. Englewood Cliffs, NJ: Prentice-Hall.
- Rayner, K. (1998). Eye movements in reading and information Processing: 20 years of research. *Psychological Bulletin*, 85, 618-660.
- Rips, L. J. (1994), *The Psychology of Proof: Deductive Reasoning in Human Thinking*, Cambridge, MA: MIT Press.
- 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.

- Siegler, R. S. & Chen, Z. (2002). Development of rules and strategies: Balancing the old and the new. *Journal of Experimental Child Psychology*, 81, 446-457.
- Siegler, R. S. (1976). Three aspects of cognitive development. *Cognitive Psychology*, 8, 481-520.
- Stanovici, K. E., Sá, W. C., & West, R. F. (2004). Individual differences in thinking, reasoning, and decision making. In J. P. Leighton & R. J. Stemberg (Eds.), *The nature of reasoning* (pp. 49-77). New York, NY: Cambridge University Press.
- Van der Maas, H., Klinkerberg, S., Straatemeier, M. (2010). Rekentuin.nl: Combinatie van oefenen en toetsen. *Examens*, 4, 10-14.
- Van Someren, M. W., Barnard, Y. F., Sandberg, J. A. C. (1994). The think aloud method: A practical guide to modelling cognitive processes. Academic Press, San Diego.
- Verbrugge, R., Mol, L. (2008). Learning to apply theory of mind. *Journal of Logic, Language and Information*, 17(4), 489-511.
- Vock, M., Preckel, F. & Holling, H. (2011). Mental abilities and school achievement: A test of a mediation hypothesis. *Intelligence*, 39, 357-369.
- Woodward, A. L. (2009). Infants' grasp of others' intentions." *Current Directions in Psychological Science* 18 (1), 53-57.
- Yenilmez, A., Sungur, S., & Tekkaya, C. (2006). Students' achievement in relation to reasoning ability, prior knowledge and gender. *Research in Science and Technological Education*, 24(1), 129-138.

Appendix 1

Learning items (L) with the solution

Appendix 2

Experimental items with the solution

Block E1-T1 items (*or*) a-d

Block E2-T1 items (*or*) a-d

Appendix 3

Experimental items with the solution

Block E1-T2 items (oo) a-d

Block E2-T2 items (oo) a-d

Appendix 4

Experimental items with the solution

Block E1-T3 items (*rr*) a-d

Block E2-T3 items (*rr*) a-d

Appendix 5

Experimental items with the solution

Block E1-T4 items (*or*) a-d

Block E2-T4 items (*or*) a-d

T4c

T4d