This is the published version of a paper published in *Journal of Mathematical Behavior*.

Citation for the original published paper (version of record):

Investigating algorithmic and creative reasoning strategies by eye tracking
*Journal of Mathematical Behavior*, 55: 100701
https://doi.org/10.1016/j.jmathb.2019.03.008

Access to the published version may require subscription.

N.B. When citing this work, cite the original published paper.

Permanent link to this version:
http://urn.kb.se/resolve?urn=urn:nbn:se:umu:diva-158187
Investigating algorithmic and creative reasoning strategies by eye tracking

Mathias Norqvist\textsuperscript{a,b,}\textsuperscript{*}, Bert Jonsson\textsuperscript{c}, Johan Lithner\textsuperscript{a,b}, Tony Qwillbard\textsuperscript{c}, Linus Holm\textsuperscript{c}

\textsuperscript{a}Department of Science and Mathematics Education, Umeå University, Sweden
\textsuperscript{b}Umeå Mathematics Education Research Centre, Umeå University, Sweden
\textsuperscript{c}Department of Psychology, Umeå University, Sweden

\textbf{ARTICLE INFO}

Keywords:
Mathematical reasoning
Cognitive proficiency
Eye tracking
Productive struggle
Solution strategies

\textbf{ABSTRACT}

Imitative teaching and learning approaches have been dominating in mathematics education. Although more creative approaches (e.g. problem-based learning) have been proposed and implemented, a main challenge of mathematics education research is to document robust links between teaching, tasks, student activities and learning. This study investigates one aspect of such links, by contrasting tasks providing algorithmic solution templates with tasks requiring students' constructions of solutions and relating this to students' learning processes and outcomes. Information about students' task solving strategies are gathered by corneal eye-tracking, which is related to subsequent post-test performances and individual variation in cognitive proficiency. Results show that students practicing by creative tasks outperform students practicing by imitative algorithmic tasks in the post-test, but also that students that perform less well on creative tasks tend to try ineffective imitative strategies.

1. Introduction

The study reported here is part of the Learning by Imitative and Creative Reasoning research program aiming at clarifying if, why, and how learning by creative mathematical reasoning can be more effective than common imitative learning strategies. In this particular study, the question of why is in focus. Earlier, qualitative studies within this research program has shown that students tend to use imitative algorithmic reasoning rather than creative mathematically founded reasoning (see below for definition of reasoning types) when solving tasks (e.g., Bergqvist, Lithner, & Sumpter, 2008; Lithner, 2003, 2008). Quantitative studies, both in mathematics education and neuro-science, have also shown that creative reasoning is more efficient in the long run than algorithmic reasoning (e.g., Jonsson, Norqvist, Liljekvist, & Lithner, 2014; Karlsson Wirebring et al., 2015; Norqvist, 2018). However, tasks that promote creative reasoning are scarce in textbooks (Jäder, Lithner, & Sidenvall, 2015; Lithner, 2004). For a broader overview of the background and research framework as well as other studies in the program, see Lithner (2017).

In order to develop broad, deep, and useful mathematical competence students need to engage in activities in which they must ‘struggle’ (in a productive sense) with important mathematics, and a delicate balance must be struck to prevent these struggles from becoming obstacles to rather than promoters of learning (Niss, 2007). It is challenging to translate this abstract idea of ‘struggle’ into the design of specific artefacts (for example, tasks) and activities useful in teaching and about the mechanisms that link such teaching to learning outcomes (Niss, 2007). The idea of a necessary ‘productive struggle’ (Hiebert & Grouws, 2007) has been studied in a number of different contexts, and researchers use terms like ‘desirable difficulties’ (Bjork & Bjork, 2011), ‘effortful retrieval’ (Pyc &
Rawson, 2009), or ‘productive failure’ (Kapur, 2010, 2014), as constructs to describe how a certain type of ‘struggle’ can improve learning. For example, Fyfe and Rittle-Johnson (2017) saw enhanced learning when children solved mathematical tasks in a “no-feedback” condition. They categorized the lack of feedback as a desirable difficulty since it triggered deeper processing and easier access to the knowledge during a delayed test. Although there is evidence that the need to struggle is crucial, we still lack an understanding of why this is. In this study we therefore investigate how task design (i.e., providing or not providing solution templates) influence how students attend information when solving mathematics tasks, and how this is related to student learning as measured by a post-test.

Although there are important insights concerning how to provide good learning opportunities (2015, Boaler, 2002; Cobb, Confrey, diSessa, Lehrer, & Schaubur, 2003; Hiebert & Grouws, 2007; NCTM, 2000; Niss, 2003; Schoenfeld, 2007; Stein, Engle, Smith, & Hughes, 2008), it is methodologically difficult to verify that the desirable learning outcomes result from teaching rather than from other variables (Niss, 2007). There is little or no transfer to competencies like problem solving ability and conceptual understanding from easier learning processes, such as imitation of given solution templates (Brousseau, 1997; Niss, 2007; Schoenfeld, 1985).

Problem solving has been a topic for research since the latter half of the 20th century, and in the 1970’s and 1980’s focus was placed on how a mathematical problem should be defined, how students solve mathematical problems, and which aspects of problem solving that could be useful to study (e.g., Schoenfeld, 1985). In more recent years, focus has shifted towards how teachers view problem solving and how teaching can be improved (e.g., Boaler, 2002; Schoenfeld, 2010, 2014; Stein et al., 2008). In regard to proposals for more effective teaching, Hiebert and Grouws (2007) concluded in a review that the state of education was far from providing a coherent and systematic knowledge base that documented robust links between teaching and learning outcomes. A similar conclusion was made by Lester and Cai (2016) in a recent review on problem posing and problem solving. They argue that problem solving research is a vast field but somewhat poorly linked to practice, and continue “although we do not know the one BEST way to teach students to be better problem solvers, research has begun to provide compelling evidence to support some methods over others”.

Problem solving has been studied extensively and continues to be of interest in mathematics education, but there is little research that compares how task designs with or without templates impact student learning, which is the focus of the present study. According to a review by Hiebert (2003), there are massive amounts of converging data showing that imitative teaching models fail to promote students’ development of central mathematical competencies effectively and instead lead mathematics students to try to follow rote learning task-solution methods (i.e., by mechanical or habitual repetition). Hiebert has concluded that students have more opportunities to learn facts and simple procedures than to engage in more complex processes, and achievement data indicate that students are indeed learning simple facts and calculation procedures but are not learning how to find solution methods by themselves or how to engage in other mathematical processes. Similar opportunities to learn have been found in a Swedish study including observations of 200 mathematics classrooms (Boesen et al., 2014). Teaching, textbooks, and assessments may promote rote learning, in the sense that algorithmic task-solution templates (e.g. worked examples, rules or standard procedures) are provided by teachers and textbooks, and many practice and test tasks can be solved by imitating such templates (Bergqvist & Lithner, 2012; Boesen et al., 2014; Jäder et al., 2015; Lithner, 2004; Shield & Dole, 2013; Stacey & Vincent, 2009; Thompson, Senk, & Johnson, 2012). The predominance of tasks with solution templates has also been discussed and criticized in the Theory of Didactical Situations in Mathematics (TDS), where Brousseau (1997) points out how a more constructive task design can improve students’ learning.

2. Research framework

2.1. The theory of didactical situations

TDS is used as a theoretical clarification of the characteristics and consequences of rote learning and as a starting point for the design of a more constructive alternative (Brousseau, 1997). First, it is used to indicate why it may be attractive (and thus prevalent) in teaching to provide algorithmic solution templates: In TDS, students’ temporary incomplete or faulty conceptions are not considered failures but are often inevitable and constitutive of knowledge formation processes. However, the teacher may try to overcome students’ obstacles by providing task-solution templates. This relieves students of the need to take responsibility for their intellectual work, and then the struggle necessary for more in-depth learning will not take place.

Secondly, the theory explains why learning by imitating algorithms is ineffective. An algorithm is broadly defined to include all pre-specified task-solving methods, such as rules and template examples. An algorithm is a sequence of executable instructions for solving a class of tasks, and it can be determined in advance. The nth transition does not depend on any circumstance that was unforeseen in the (n-1)th transition – it does not depend on new information, new decisions, interpretations, or thus on any meaning that could be attributed to the transitions. Therefore, the execution of an algorithm has high reliability and speed, which is a strength when the purpose is only to solve a task. However, if the purpose is to learn, an algorithm executed without considering its meaning may lead to rote learning. It is the domination of algorithmic solution templates in mathematics teaching and learning, not the algorithms themselves, that is problematic. However, algorithms are a fundamental and crucial part of mathematics. Fan and Bokhove (2014) have concluded in their literature survey that “learning of algorithms has suffered from an alleged dichotomy between procedures and understanding” (p. 481) but also that “the majority of more recent research seems to indicate that products and processes, procedures and understanding, go hand in hand” (p. 484).

Thirdly, the aim of TDS is the design of situations that allow for the construction of knowledge by the learner (as an alternative to imitation). One central aspect is the devolution of problems: Students must take responsibility for a part of the problem-solving process. The teacher’s task is to arrange a suitable didactic situation in the form of a problem in such a way that if students solve it, then the
students will obtain the desired target knowledge. From the point when the students accept the problem as their own to the moment when they produce an answer, the teacher refrains from interfering and suggesting how to solve the task. The teacher must therefore arrange the devolution of a good problem rather than describe what the students are supposed to do. This does not imply that the teacher is more passive or has a less important role than that in the ‘solution-template providing’ approach. Designing a good problem for devolution is usually much more difficult than designing imitative tasks, and it places higher demands on teacher interaction. In this study, tasks that are potentially creative in nature are contrasted with tasks where a solution template is given. According to TDS, there should be differences between these two types of task, both regarding how students direct their attention to information that can provide deeper mathematical insight, and in learning outcomes.

2.2. Algorithmic and creative reasoning

A series of studies resulting in a research framework (Lithner, 2008) have suggested that a key factor affecting learning outcomes is whether students engage in algorithmic or creative reasoning. Reasoning is here defined as the line of thought adopted to produce assertions and reach conclusions in task solving. It is not necessarily based on formal logic; thus, it is not restricted to proof and may even be simple, incorrect and/or superficial, as long as some sensible reasons (as perceived by the reasoner) support it.

Algorithmic reasoning (AR) consists of an attempt to solve a task by applying a given or recalled algorithm. Examples include following a memorized procedure of finding the line through two points or imitating an example given by the teacher of how to multiply two three-digit numbers and applying it to two other numbers.

Opportunities for students to create knowledge in line with TDS have been found to be rare in teaching, textbooks and tests. When it is applied, students are able to make better progress with Creative Mathematically founded Reasoning (CMR). Empirical studies of the distinctions between AR and students’ own constructions of solutions have defined this reasoning type as fulfilling three criteria: (a) Creativity: the learner creates a reasoning sequence not experienced previously, or re-creates a forgotten one (Silver, 1997). (b) Plausibility: there are predictive arguments supporting the strategy choice and arguments for verification, explaining why the strategy implementation and conclusions are true or plausible (Lithner, 2008; Pólya, 1954). (c) Anchoring: the arguments are anchored in the intrinsic mathematical properties of the components of the reasoning (Lithner, 2008). A literature review has revealed two main uses of mathematical ‘creativity’ (Sriraman, Haavold, & Lee, 2013): the extraordinary creativity of geniuses and the everyday creativity that “can be fostered broadly in the general school population” (Silver, 1997, p. 75). The latter meaning is used here, i.e., the creation of task solutions (or the re-creation of forgotten ones) that are original to the individual who creates them.

CMR has been observed in think-aloud studies where students solve tasks individually (e.g., Bergqvist et al., 2008; Lithner, 2003), as well as in mathematical discussions between students (e.g., Granberg & Olsson, 2015; Hershkowitz, Tabach, & Dreyfus, 2017). Solving a task using CMR is largely similar to what many others have written in terms of (non-routine) problem solving (NCTM, 2000). When a “problem” is defined as a task for which students have no access to a solution method from the start, there are usually in the literature various additional requirements related to potential struggle, for example, that a problem should be challenging (Schoenfeld, 1985) or require exploration (Niss, 2003). A task requiring CMR does not have to meet criteria such as being challenging, requiring exploration or invoking modelling, but it has to incorporate parts where construction of (to the student) new mathematical reasoning is required. Thus, this study will not focus on problem-solving skills in general, but on the particular distinction between imitation and creation of solution methods which is one central aspect of problem solving (Brousseau, 1997; Schoenfeld, 1985). Additionally, pilot studies and think-aloud studies (Bergqvist et al., 2008; Liljekvist, 2014; Lithner, 2003) have shown that tasks that provide solution templates are mainly solved by AR, since the solution template provides a safe and time-efficient method. Henceforth, tasks that provide solution templates will be denoted AR tasks, since it is likely that students will attempt AR to solve them. If no solution template is available in the task, and it is not solvable by complete standard methods known to the student, then the student has to use CMR at least in some part to solve it, and therefore these tasks will henceforth be denoted CMR tasks.

2.3. Comparing AR and CMR empirically

Within the Learning by Imitative and Creative Reasoning research program several studies have investigated the distinction between CMR and AR. In a sample of upper secondary school students (n=130) Jonsson et al. (2014) found that practicing CMR was superior to practicing AR at a follow-up test one week later (see below for examples of AR tasks with solution templates and CMR tasks without information about how to solve them). The study also showed that cognitive proficiency was significantly associated with test performance, but more so for participants practicing AR. In other words, it was the students with the lowest cognitive proficiency (within the sample) that had most to gain by CMR practice compared to AR practice. Although previous research has shown similar results (Koichu, Berman, & Moore, 2007), this finding contradicts the common belief that tasks requiring creative reasoning are more suitable for high-performing students. In a similar study with upper secondary students (n=104), Norqvist (2018) added written explanations developed by experienced teachers to the AR tasks (but not to the CMR tasks), explaining why the given solution procedures worked. However, Norqvist (2018) found no gain in post-test performance for participants practicing AR relative to those practicing CMR. In Jonsson et al. (2014) it was argued that the superior CMR performance is in line with the effortful retrieval hypothesis, that more effortful encoding facilitates later performance (Pyc & Rawson, 2009; van den Broek, Takashima, Segers, Fernandez, & Verhoeven, 2013), and with the argument that suitable struggling with important mathematics is necessary for subsequent performance (e.g., Hiebert & Grouws, 2007; Niss, 2007). However, an effect of the similarities between practice tests and tests tasks, denoted as transfer inappropriate processing, could not be ruled out. To pursue this argument Jonsson, Kulaksiz, and
Lithner, (2016) contrasted CMR and AR tasks with respect to effortful struggle and transfer appropriate processing. Although there was an effect of transfer appropriate processing the main cause of the difference between AR and CMR was found to be the mathematical struggle. Notably, it was not the struggle per se that was important, but rather whether the amount of struggle faced with during practice leads to correct answer or not. In Karlsson Wirebring et al. (2015) the AR/CMR distinction was further investigated using a brain imaging technique, denoted as functional Magnetic Resonance Imaging (fMRI). Karlsson Wirebring et al. (2015) showed that CMR practice lead to reduced cognitive load at post tests and superior test performances. These findings were evident from greater brain activation in precentral cortex (Brodmann area 6) for participants that had been practicing by AR relative those that had practiced by CMR. This area is associated with working memory related tasks (Desposito et al., 1995), such as tasks with high cognitive load. In line with the assumption that AR practice lead to a focus on memorising and recalling the provided formulas (2017, Lithner, 2008), Karlsson Wirebring et al. (2015) also discovered that AR, to a greater extent than CMR, engaged brain regions that are known to be important for operations requiring access to verbal memory of facts (the left angular gyrus) at the post-test. This was interpreted as attempts to recall, from long-term memory, the formulas which the participants solving AR tasks repeatedly were exposed to during practice.

In summary, behavioral and fMRI data have shown that CMR practice is superior to AR (Jonsson et al., 2014, 2016; Karlsson Wirebring et al., 2015). AR tasks seem to invite students to use the provided algorithms without having to regard the mathematical meaning of the tasks and their solutions, while CMR tasks seem to invite/force students to do this (Jonsson et al., 2016). The latter will during practice be more cognitively demanding than imitating a provided algorithm, and therefore cognitive abilities are taken into account in this study.

2.4. Cognitive requirements and attentional strategies

When investigating educational interventions that are high in terms of their cognitive requirements (e.g., mathematical task solving), we argue that it is not only important to consider the didactical context (e.g., the specific teaching situation and how tasks are designed), but also individual variations in cognitive abilities. Working memory (WM) and non-verbal problem solving has repeatedly been identified as a significant predictor in school performance (Alloway, Gathercole, Kirkwood, & Elliott, 2009; Andersson & Lyxell, 2007; Gathercole & Pickering, 2000; Hitch, Towsse, & Hutton, 2001; Primi, Ferrão, & Almeida, 2010). In mathematical task solving, WM is responsible for the online manipulation of transient information, and the resulting transfer of information to long-term storage. The idea of WM was developed from the concept of short-term memory and includes an attentional executive control system that controls three separable, but interacting, subsystems: the phonological loop, the visuo-spatial sketchpad, and the episodic buffer (Baddeley, 2000). Through these three subsystems, WM both feeds information into and retrieves information from long-term memory.

Non-verbal reasoning reflects humans’ ability to flexibly adapt their thinking to new problems and situations and is regarded as relatively independent of education. It is viewed as a measure of fluid intelligence which in turn is a prominent factor of general intelligence and an important predictor of mathematical achievement (Primi et al., 2010).

WM and nonverbal reasoning abilities are cognitive proficiencies that have a high impact on school achievements that also varies considerably between individuals. An assumption that seems common is that cognitive less proficient students are best helped if provided with algorithmic support (Boesen, 2006). In addition, studies have found that variance shared between working memory capacity and fluid intelligence is related to the ability to control attention (Unsworth & Engle, 2005), and that cognitive less proficient students tend to experience problems while simultaneously processing information and directing attention to the, for the task, relevant information (e.g., de Fockert, Rees, Frith, & Lavie, 2001; Kane, Bleckley, Conway, & Engle, 2001). For a cognitively less proficient student, a specific task will be cognitively more challenging than for a cognitively more proficient student. Instructional design can be used to reduce the cognitive load during the learning phase, in particular, the working memory load (e.g., Sweller, van Merrienboer, & Paas, 1998). In the present study no attempt was made to reduce the cognitive load during practice in either of the learning conditions. However, to ensure that cognition did not confound the results a matched pair design was used blocking cognitive proficiency (i.e., working memory capacity and non-verbal reasoning), gender, and grade.

2.5. Eye tracking and mathematics

Eye tracking, i.e., measuring the point of gaze in relation to a visual stimulus, is an established way to determine what parts of the stimulus are being attended to, and for how long (Duchowski, 2003; Rehder & Hoffman, 2005). Although it is well known that attention, to some extent, can dissociate from gaze during intentional covert attention (Posner, 1980) eye movements principally follow shifts in attention (Kowler, Anderson, Dosher, & Blaser, 1995) and eye fixations are highly coupled to the locus of attention for all but the simplest stimuli (Deubel & Schneider, 1996). We therefore argue, that visual attention in terms of eye fixations or dwell time can provide a window to the viewer’s cognitive processes (Yarbus, 1967) and their understanding of the task at hand (Jarodzka, Scheiter, Gerjets, & van Gog, 2010). Studies in mathematics education have shown that eye-tracking can provide information about the attentional processes involved in mathematical task solving. Schindler, Lilienthal, Chadalavaida, and Ogren, (2016) argued that although video recordings of students’ task solving revealed in which order solution steps occurred, eye-tracking offers a more fine-grained access to students’ task solving process. Susac, Bubic, Kaponja, Planinic, and Palmovic, (2014) conclude that experts are better at directing attention to relevant information (had less fixations) than novices, and that this difference increases with increasing task difficulty. This was also discovered by Lin and Lin (2014) who suggested that the higher number of fixations among novices could be due to difficulties in relating complex visual stimuli to memorized referents. Experts have also been shown to spend
more time on logical structures when reading mathematical proofs than novices, who often focus on surface features (Inglis & Alcock, 2012). Susac et al. (2014) also argued that eye-tracking can be a more objective and reliable measure of metacognition than reports from the participants. Eye-tracking has also been found to be useful to establish which solution strategies children use (Obersteiner & Tumpek, 2016). This could be very important, especially among children that are having trouble verbalizing their strategies. In a study by Hegarty, Mayer, and Monk, (1995), “short cut” approaches (using keywords and relational terms such as “more” or “less”) and meaningful approaches (translating the mathematical task into a mental model of what is described) gave rise to different eye fixation behaviors. Participants using shortcut approaches were less successful, and their eye tracking behaviors were characterized by repeated reexamination of numbers and relational terms and disregard of words that were more informative for the task at hand. This finding is aligned with the reasoning put forward by Lithner (2004), 2008) and Jonsson et al. (2014), that focusing on more superficial information does not facilitate a conceptual understanding of the task solution method. In the present study, we investigate what type of information the participants are focusing on during practice of either AR or CMR tasks. Each task layout is defined according to five different information areas, denoted as “areas of interest”. Depending on practice task type (AR or CMR) each area of interest contains more or less valuable information, see Fig. 1 for examples.

3. Aim and hypotheses

From the background presented above, it is relatively clear that learning by CMR can be more efficient than learning by common imitative algorithmic methods, at least in some contexts. However, in order for such information to be useful also in other contexts one important type of information concerns insights into why different types of learning strategies may be more or less functional. Therefore, the purpose of the present study is to investigate if and how students’ learning strategies affect learning by relating theoretical assumptions about learning strategies (i.e., that tasks with or without solution templates will invite different types of reasoning) to time spent gazing in areas of different task information (e.g., the illustration will be important for constructing a solution when no template is given) and comparing this to post-test performance (see below for details). Unsuccessful solvers have a harder time finding the task relevant information (Lin & Lin, 2014). This made it relevant to investigate the association between dwell time on task-relevant information and post-test performance. Individuals with lower working memory capacity do, to a greater extent, experience problems directing their attention to the relevant information (e.g., de Fockert et al., 2001; Kane et al., 2001). Measurements of cognitive proficiency (e.g., working memory and non-verbal problem-solving ability) were therefore used as a control variable when investigating the associations between gaze and test performance. Although the main focus was on learning strategies, index by eye tracking, the initial hypothesis 1 was on replicating previous studies which have shown that practicing with CMR is superior to practicing with AR tasks on later tests. We therefore tested three hypotheses:

1 Based on Jonsson et al. (2014), Karlsson Wirebring et al. (2015), and Norqvist (2018) it was expected that participants that practice by CMR tasks would outperform participants that practice by AR tasks on a subsequent test session where the practiced solution methods were tested (see the section on test tasks below).

2 Participants that practiced on AR tasks were expected to focus more on areas with information about how to carry out the algorithm and disregard information useful for understanding why the algorithm works (e.g., an illustration helping the student to see the relation between the task and the algorithm). Participants that practiced on CMR tasks were expected to focus more on the information useful for seeing and using such relations since this would be needed to construct a correct solution method (Brousseau, 1997; Hegarty et al., 1995; Jonsson et al., 2014; Lithner, 2008).

3 While controlling for cognitive proficiency, positive associations between dwell time, in areas that contain task relevant information, and post-test performance were expected for both groups of participants.

Fig. 1. Example of the two task-types, AR (a) and CMR (b) and the areas of information (1–5).

Note. The dotted squares defining the areas of interest are for illustrative purpose and were not visible for the participants.
4. Method

4.1. Participants

Fifty participants, with a mean age of 23.0 years ($SD = 3.2$), were recruited from an upper secondary school and from a university in Sweden. The upper-secondary students were recruited from the natural-science program that focuses on mathematics and science. The university students came from the psychology program, which requires high grades from upper-secondary school. The participants can therefore be regarded as mid to high achievers in mathematics. Written informed consent was obtained in accordance with the Declaration of Helsinki, and the study was approved by the Regional Ethical Review Board, Sweden. After adjusting for absence during one of the sessions, the data of 48 participants (25 CMR and 23 AR) were analyzed.

4.2. Materials and design

Eleven tasks, novel to the students, that invited learning through AR or CMR during practice were used, and performance as a function of practice was measured. The tasks were a subset and adapted version of Jonsson et al. (2014) tasks. Out of 11 tasks, 10 were used in the analyses. Task 11 was found to be too difficult irrespectively of practice conditions. The study was conducted in three sessions: a cognitive test session, a mathematics practice session with eye-tracking, and a test session. During the cognitive session, the participants took two cognitive tests. In a matched pair design, participants were blocked according to cognitive ability, grade and gender and randomly assigned to either AR or CMR conditions. The design ensured that the two groups (denoted AR and CMR groups) were as similar as possible regarding cognitive abilities and mathematical proficiency and that there was approximately the same number of men and women in each group. The design explicitly controlled for these potential confounders and thus provided a strong case that any obtained differences in performance would be caused by the manipulation. All participants worked individually during the practice session, and were subsequently tested individually. A researcher was present in the lab during both the practice and test sessions to monitor the procedure. During practice and test no assistance was provided, except for answers regarding the use of the computer. In the following sections the cognitive measures, practice tasks, and test tasks are outlined and described.

4.3. Cognitive measures

The participants were initially measured on two cognitive variables: Raven's Advanced Progressive Matrices (Raven, 1991) and Operation Span (Unsworth, Heitz, Schrock, & Engle, 2005). Operation span is a measure of working memory capacity with a good test-retest reliability and high internal consistency (e.g., Conway, Cowan, Bunting, Therriault, & Minkoff, 2002; Engle, Tuholski, Laughlin, & Conway, 1999; Klein & Fiss, 1999) and correspond well with other measures of working memory capacity and higher order cognitive tasks (Conway et al., 2002; Unsworth & Engle, 2005). Operation span is computer-based and responses are entered using a computer mouse. The task is divided into processing and storing information simultaneously followed by retrieval of the stored information. In sets of three to seven, mathematical true or false task is presented (e.g., $3 + 7 = 5$) and after each task a single letter is presented. After each set the participants are asked to recall the letters in correct order by selecting the correct letters (among distractors). In total, the participants are asked to solve 75 mathematics tasks and recall 75 letters in correct position. The dependent measure used in the analysis of working memory capacity was the total of letters in correct position (see Unsworth and Engle (2005) for a detailed description). Raven's advanced progressive matrices is a well-known standardized test of non-verbal reasoning. Raven's matrices comprise 36 items that are presented in ascending order of difficulty. Each item consists of a $3 \times 3$ grid with eight symbols and one empty cell. The task is to complete the pattern by selecting the missing symbol among eight alternatives. In the present study, we used 18 items (every other), thus maintaining the ascending order of difficulty. The participants were allowed 20 min and the total score was collected. The order of Operation span and Raven's matrices were counterbalanced within each group. Several studies have shown that working memory capacity and its executive functions are intimately related to arithmetic performances (e.g., Cragg & Gilmore, 2014). This is also the case for Raven's matrices (e.g., Primi et al., 2010). The correlation between operation span and Raven's matrices was found to be .42 ($p = .003$). Both operation span and Raven's matrices scores were transformed to $z$-scores and merged to a composite score, denoted as cognitive proficiency index (CPI), and used in the matching procedure described above. The CPI was also used to make comparisons within our sample.

4.4. Practice tasks

The target knowledge for both AR and CMR was solution methods for 10 different sets of mathematical tasks. For instance, to find out how many matches that are needed to form a row with a specific number of squares (see Fig. 1).

Participants in the AR group were provided with 10 numerical subtasks for each of the 10 tasks sets with a time limit of totally 5 min per task set. For each individual subtask, the solution method was provided, as well as an example of how to apply it (Fig. 1a). If the participants were able to complete more than ten subtasks within the 5-minute time limit, the computer re-sampled from the previous subtasks.

Participants in the CMR group were also given 10 numerical subtasks for each of the 10 task sets with a time limit of totally 5 min per task set, however, in contrast to AR there were no algorithm or other guidance on how to solve the tasks (Fig. 1b). A CMR task takes in general more time to complete than the corresponding AR task, so in most cases the participants in the CMR group would not
complete all 10 tasks in every task set during practice. In the unlikely event that a participant would be able to complete more than 10 subtasks in each task set within the 5-minute time limit, the computer re-sampled from the previous tasks. The mean solution time for each sub-task differed between the AR and CMR group (27 s and 56 s respectively).

The imitative AR approach is similar to what usually is provided by textbooks and/or by teachers, while the CMR approach can be seen as an alternative to enhance learning through students’ own mathematical reasoning. For both AR and CMR practice tasks, each image presenting the tasks were divided into five areas of interest. This division seen in Fig. 1a-b was not visible to the participants. Each area was denoted according to its specific information; (1) illustration, (2) description, (3) formula, (4) example, and (5) question. Examples and formulas were only available for AR participants. For CMR participants those areas are replaced with text containing no information about how to solve the task.

4.5. Test tasks

The test tasks were identical for both AR and CMR participants. Each of the 10 task sets consisted of three subtasks, denoted as test task I-III (Fig. 2a-b). In test task I the participants had a time limit of 30 s and were asked to write down a formula associated to the specific task solution method (Fig. 2a). This subtask was novel for all participants, irrespectively of group. It was (from piloting data) judged that 30 s was sufficient to read the task, recall the answer from memory and write it down but not to (re)construct it. One may note that the CMR group may not have considered any formula during practice, since all practice tasks were possible to solve without representing the solution method as a formula. The AR group had, on the other hand, been shown a formula in all practice subtasks. In Test task II the participants had 30 s to complete a numerical task (Fig. 2b). It was judged that 30 s is too short to (re)construct a solution method, but enough time to recall and apply a memorized solution (e.g., a formula or a principle for finding the number of matches). In test task III the time limit was set to 300 s, providing enough time to also (re)construct a solution method if forgotten (Fig. 2b).

4.6. Eye tracking equipment

Eye movements were recorded with a desk mounted Eyelink 1000 sampling at 500 Hz. A chin-rest was used to impose constant viewing distance of about 75 cm from the screen and stabilize participants’ heads. Eye movements were recorded monocularly, typically using the right eye. Eye-fixations were defined as non-blink inter-saccadic intervals based on the manufacturer’s default settings. The saccade definitions consisted of an amplitude change of 0.15°, velocity > 30°/ms and an acceleration threshold of 8000°/ms². Only fixations thus defined with durations longer than 50 ms were considered for further analysis. In-plane gaze spatial resolution was about one degree across participants.

4.7. Data analyses, an overview

To capture the unique task solution strategies characteristics of AR in comparison to CMR during practice, we therefore decided to restrict the analyses to the first three subtasks in each task-set. This was done since it is likely that students that construct their own solution method (i.e., the participants in the CMR group), after confirming it in a few additional tasks, will solve the tasks without new construction, thus in an AR manner. There are also clear indications that students that meet tasks that include a solution template or a formula (as in AR tasks), tend to use that template to solve the task. The CMR tasks do not include a solution template, and it is unlikely that the students already know the solution for the particular tasks used in the present study. Hence, it is likely that the student will use CMR to solve the task. To confirm that the AR and CMR groups were equal in cognitive proficiency after removal
of participants that did not complete all sessions, an independent t-test was conducted. The analysis revealed that the cognitive proficiency of the AR group ($M = .16, SD = .68$) did not significantly differ from the CMR group ($M = -.17, SD = .97$), $t(46) = 1.37$, $p = .180$, two-tailed.

To evaluate the main findings from Jonsson et al. (2014), whether CMR practice is superior to AR on a later test, independent t-tests of practice and test scores were conducted (Hypothesis one).

Since the time spent gazing in each area of interest is a relevant measure of how the participants regard and process the information, we chose to use dwell time as the measure of gaze in our analyses. To study where participants focused and whether this focus was dependent on practice group (Hypothesis two) a hierarchical cluster analysis was conducted. The hierarchical clustering extracts information (i.e., proportional dwell time) from all areas of interest and utilizes this to create clusters. Using Block distance and Ward’s method (Ward, 1963), variance within clusters are minimized and variance between cluster are maximized and through an iterative process amalgamated into clusters of increasing dissimilarities. As a part of the result section below, the clusters will be described in detail.

Finally, to reveal potential associations between post-test scores and gaze on information, relevant for task solving, while controlling for cognitive proficiency (Hypothesis three), partial correlation analyses were conducted. Effect size measures were used to estimate the magnitude of obtained group differences. Cohen’s $d$ was used when sample sizes and standard deviations were similar in the groups. For the sub-cluster analyses, which encompassed groups with fewer participants, Hedge’s $g$ was calculated to get an estimate of effect sizes. An effect size of about 0.2 is interpreted as small, 0.5 as medium, and 0.8 as large (Cohen, 1988). All statistical analyses were conducted using the Statistical Package for the Social Sciences, version 24 (SPSS 24).

5. Results

Although the AR-group did perform better during practice, $t(32) = 5.56, p < .001, d = 1.58$, they were outperformed by the CMR-group during the test, $t(46) = −2.44, p = .019, d = .71$ (see Fig. 3). The results from Jonsson et al. (2014) were therefore confirmed (Hypothesis 1).

To control if any difference in gaze could be referred to the two practice groups (Hypothesis 2), proportional dwell time in the different areas of interest (i.e., illustration, description, formula, example, and question) were entered into a hierarchical cluster analysis. Three clusters (A, B, and C) emerged from this analysis, and a closer look on the three clusters showed that there was one cluster with AR-students only (cluster A), one cluster with mainly AR-students (cluster B), and one cluster with only CMR-students (cluster C) (see Fig. 4).

The main focus in cluster A was on the formula (Fig. 1), in cluster B on the example and in cluster C on the illustration. Clusters B was divided into two sub-clusters, AR-B ($N = 11$) comprised participants that practiced with AR-tasks and CMR-B ($N = 6$) comprised participants that practiced with CMR-tasks. Cluster A and Cluster C included only participants from one practice condition (AR and CMR respectively) and are therefore denoted AR-A ($N = 12$) and CMR-C ($N = 19$). In addition to the main focus of clusters A, B and C, the specific differences in focus between these clusters can be seen in Fig. 5. The illustration draws more attention from the participants in both CMR sub-clusters, and is essentially ignored by the AR sub-clusters. The descriptions are not focused on by any group. In the AR tasks, the formula and example each contain complete solution templates by themselves and AR-A and AR-B are
distinguished between which of the two that is preferred. Also, CMR-B directs most attention to the example, but this does not include any useful information for the CMR-group. The question is focused on to a similar degree, and to a relatively large extent, by all groups.

Finally, an analysis of whether the dwell time in areas containing useful information would be correlated to practice and test scores was conducted. The rationale for these analyses is that gaze on task relevant information should be positively associated with test-task performances. However, an initial analysis revealed significant ($p < .01$) and positive correlations between cognitive proficiency and practice score, and between cognitive proficiency and test score; $r(48) = .60$ and $r(48) = .43$, respectively. To control for cognitive proficiency, we therefore conducted a partial correlation analysis between the proportion of dwell time within each of the five areas of interest and the test score, with cognitive proficiency as a control variable. The analyses revealed no significant correlations between dwell time and test score for the AR-group. For the CMR-group two significant correlations were
found. First, a positive correlation ($r(25) = .53, p = .008$) between test score and dwell time within the illustration. Second, a negative correlation ($r(25) = −.65, p = .001$) between the test score and dwell time within the example. This negative correlation was driven by the participants in CMR-B who had their main focus on the example. With respect to differences between CMR-B and CMR-C analyses of effects sizes revealed that: (a) CMR-B had substantially longer total dwell time than the participants in CMR-C ($g = 1.39$), (b) CMR-C out-performed CMR-B on both practice and test ($g = 1.15$ and $g = 1.33$, respectively), and (c) CMR-B had lower cognitive proficiency than participants in CMR-C ($g = 0.84$) (Fig. 6).

In summary, the results show the following:

1) Hypothesis 1 was confirmed, the CMR group outperformed the AR group in the post test.
2) Hypothesis 2 was partly confirmed. With regard to the information available in the task, participants did not exclusively focus on information useful for constructing a solution. Instead, the cluster analysis divided the participants into three clusters with different main foci (formula, example, or illustration), yielding four sub-clusters when considering practice condition. AR-A, AR-B, and CMR-C acted in line with the hypothesis, mainly focusing areas judged to be useful for solving the task: The two former sub-clusters focused the formula and the example respectively, that both contain algorithmic templates, and the latter sub-cluster paid much attention to the illustration, which is reasonable to do in order to construct a solution. This was not the case for CMR-B, who mainly focused on the seemingly non-useful information in the CMR-example and whose participants had lower cognitive proficiency and longer dwell time than participants in CMR-C. One may note that the illustration was not regarded by the AR participants.
3) For CMR participants there was a positive correlation between their focus on the illustration and the test score, and a negative correlation between their focus on the un-useful example and test score.

6. Discussion

This study aimed to increase our understanding of the learning strategies used by students when practicing by either AR or CMR tasks. Previous studies have shown that CMR practice yields better test scores than AR practice (Jonsson et al., 2014; Karlsson Wirebring et al., 2015; Norqvist, 2018). These previous findings were replicated in the present study. Hence the CMR participants outperformed the AR participants, which confirmed the first hypothesis stating that practicing with CMR tasks is more efficient than practice AR for later test performance (in the setting of this experiment).

6.1. Potential reasons behind learning strategies for AR-tasks

The second hypothesis concerned the difference in eye-fixation between the two practice groups, expecting the AR group to focus on the formula and/or the example, while the CMR group was expected to fixate more on the illustration. The general assumption was that irrespectively of practicing CMR or AR the participants will use the information that is most effective for solving the tasks. In line with hypothesis two, we confirmed that the behavior of the AR participants was, to a large extent, characterized by attending

Fig. 6. Practice and test scores for the four sub-clusters.

Note. Error bars represent one standard error of the mean.
information that is effective in the short term (i.e., applying the formula and/or the example). This information bias was in addition characterized by almost no dwell time on the illustration, while the CMR participants attended to the illustration as indexed by a higher dwell time. In Karlsson Wirebring et al. (2015) it was found that brain activity on later test for those that had practice with AR tasks, was restricted to significant angular gyrus activation. This was interpreted as attempts to retrieve information that they repeatedly had been exposed to, namely the formulas/examples. The findings of the present study agree with that conclusion. Hence practicing with AR tasks directs the participant’s attention to the formula.

A question to ask is why most AR participants focused more on the formulas, examples and the questions and choose to disregard the illustration. In economic terms it is more efficient to use the given formula and the example as they provide enough information to solve the task. Hence, the finding that AR participants disregarded the illustration seems to indicate that they used the formula without analyzing the basic properties of the pattern displayed in the figures. This is in line with the arguments by Brousseau (1997) that algorithms are designed to be efficient, give the correct answer (if applied correctly) and to avoid the need to consider the meaning of the mathematical properties of the task. Thus, this strategy may lead to superficial rote learning. AR tasks seem to dominate common textbooks (Boesen et al., 2014; Jäder et al., 2015; Lithner, 2004; Shield & Dole, 2013; Stacey & Vincent, 2009; Thompson et al., 2012). The results of this study indicate that if we as teachers hope that students will do more than mimic templates, e.g. reflect over the mathematical meaning of such tasks and solutions, this hope may be in vain.

6.2. Potential reasons behind learning strategies for CMR-tasks

The cluster analysis revealed two distinct CMR sub-clusters instead of one as was expected. In CMR-C the participants focused mainly on the illustration, and in CMR-B the participants focused mainly on the example. Since the examples carried no useful information for the CMR participants, the obtained CMR-B sub-cluster was somewhat surprising.

Although we didn’t manipulate the amount of information, the result in the present study indicates that providing information that is redundant- or even of questionable value, can potentially hamper task-solving. A possible reason could be that individuals with lower working memory capacity have more difficulties in sorting out the redundant information (e.g., de Fockert et al., 2001; Kane et al., 2001). Both the task-relevant and task-redundant information has to be processed on-line which in turn requires executive attention and thus a selection of the task-relevant information. Hence hampering the cognitively weaker students, for which the search process is more effortful, and potentially causes them to use task-redundant information. In the present study this behavior was indicated by participants in CMR-B. Another reason for the focus on the example by the CMR-B cluster group could be a lack in problem-solving skills (i.e., heuristics, meta-cognition, beliefs, and resources, see Schoenfeld, 1985) which leads students to seek for familiar AR-information. That the participants in the CMR-C sub-cluster pays more attention to the illustration can be an indication of creative reasoning that starts off with the illustration. In qualitative pilot studies the illustration has been found to be used as a basis for reasoning about the relations between this type of tasks and the solution method. Also, in more general non-routine problem solving, illustrations with mathematical information are used to construct solutions (Schoenfeld, 1985). However, the illustration itself does not tell the student how to solve the task. This requires construction of a solution method and this construction is what Brousseau (1997) and TDS argues for. Trying to construct a solution with less regard to task-relevant information, in this case the illustration, will not be efficient for learning; as can be seen from the performance of the CMR-B group.

6.3. Relations between learning strategies and test scores

There should, according to the third hypothesis, be a positive association between the time spent gazing in areas that contain task-relevant information and post-test performance. For the AR-group, the analyses did not reveal any positive associations between dwell time within any of the areas of information and post-test performance. There are, together with the question, enough information to solve the task in both the formula or the example separately for the AR-group. Thus, which one of these two areas a student chooses to put her main focus is of less importance.

For the CMR-group there were correlations between the dwell time within some areas of interest and the post-test score. First, there was a positive correlation connected to dwell time on the illustration. Time spent gazing at the illustration could indicate that the participant tries to understand how the patterns or geometrical figures are constructed (i.e., the mathematical properties of the task). Second, there was a negative correlation between dwell time on the example, which carried no useful information, and the post-test score. This finding is driven by the CMR-B participants’ disproportional focus on the example, which also was characterized by their inferior performance on both practice and test tasks. Participants that, as pointed out above, also scored lower on the cognitive tests and potentially lack problem solving skills.

Ongoing qualitative studies indicate that students sometimes use non-effective strategies when encountering problems. For example, some students wrongly try to construct solution methods based on a single example, either by a relation found in an illustration or by calculating backwards. In the match-stick task above (Fig. 1) students have been observed to calculate an average number of match-sticks per square (i.e., 13/4 = 3,25) and use this to try to calculate the number of matches needed for 50 squares (i.e., 50·3,25 = 162,5). Such an approach could possibly have been used by the CMR-B group, which would be in line with both their gaze and their low post-test results. The possibility to control their solution method by expanding the illustration with two more squares and calculate and check with the new figure does not seem to be a heuristic that these students are used to utilize. Similar erroneous solution methods have also been observed in other tasks.

Considering the limitations of this study, additional or different background variables could, of course, have influenced the matching of the practice groups. We did, however, decide that cognition, gender, and mathematics grade would be sufficient to make
a reliable matching. The slight (but not significant) difference in CPI could potentially have caused an underestimation of the group difference at the test, since the AR-group on average showed a slightly higher CPI. The small sample size, especially in the sub-clusters, is also a limitation. Because of this we did not conduct any inferential statistics, but report effect size as a measure of the difference between the sub-clusters.

7. Conclusion

Studies have shown that in order to develop a deeper understanding in mathematics, students need to struggle with mathematical properties and, hence, to meet tasks that are designed with this in mind (e.g., Fyfe & Rittle-Johnson, 2017; Hiebert, 2003; Jonsson et al., 2014; Niss, 2003; Schoenfeld, 1985). There are also arguments for the in-effectiveness of imitative procedural tasks regarding the development of mathematical understanding (e.g., Brousseau, 1997; Hiebert, 2003). This study gives us an insight into why procedural tasks are less effective than creative tasks. Students that are exposed to given algorithms tend to disregard task elements that promote understanding of the solution method. In the present study students practicing by AR-tasks generally disregard the illustration, which is important to build deeper understanding of the algorithm (i.e., why the algorithm works). Students that practice by CMR-tasks generally give more attention to the illustration. However, eye-tracking has provided quantitative evidence that not all students behave as can be expected based on the task design. A few students that practice by CMR-tasks behave as AR-participants and this seems crucial for their low performance both in the practice and test situation. This supports the conclusion that teachers have an important role to play in the classroom, not only as provider of well-designed non-routine problems but also as a supporter of students’ own attempts to construct mathematical knowledge and become proficient mathematical problem solvers.

Declarations of interest

None.

Acknowledgements

This study was supported by the Marcus and Amalia Wallenberg Foundation and Umeå University, Sweden.

References


