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

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

Learning mathematics through algorithmic and creative reasoning.
*Journal of Mathematical Behavior*, (36): 20-32
http://dx.doi.org/10.1016/j.jmathb.2014.08.003

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-95773
Learning mathematics through algorithmic and creative reasoning

Bert Jonsson\textsuperscript{a,}\textsuperscript{*}, Mathias Norqvist\textsuperscript{b,d}, Yvonne Liljekvist\textsuperscript{e,f}, Johan Lithner\textsuperscript{c,d}

\textsuperscript{a} Department of Psychology, Umeå University, Sweden
\textsuperscript{b} Department of Mathematics and Mathematical Statistics, Umeå University, Sweden
\textsuperscript{c} Department of Science and Mathematics Education, Umeå University, Sweden
\textsuperscript{d} Umeå Mathematics Education Research Centre, Umeå University, Sweden
\textsuperscript{e} Department of Mathematics and Computer Science, Karlstad University, Sweden
\textsuperscript{f} The Centre of Science, Mathematics and Engineering Education Research, Karlstad University, Sweden

\section*{A R T I C L E   I N F O}

\textbf{Keywords:}
Mathematical reasoning
Reasoning
Cognitive proficiency
Memory retrieval

\section*{A B S T R A C T}

There are extensive concerns pertaining to the idea that students do not develop sufficient mathematical competence. This problem is at least partially related to the teaching of procedure-based learning. Although better teaching methods are proposed, there are very limited research insights as to why some methods work better than others, and the conditions under which these methods are applied. The present paper evaluates a model based on students’ own creation of knowledge, denoted creative mathematically founded reasoning (CMR), and compare this to a procedure-based model of teaching that is similar to what is commonly found in schools, denoted algorithmic reasoning (AR). In the present study, CMR was found to outperform AR. It was also found cognitive proficiency was significantly associated to test task performance. However the analysis also showed that the effect was more pronounced for the AR group.

\textcopyright\ 2014 The Authors. Published by Elsevier Inc. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/3.0/).

\section*{1. Introduction}

The overarching goal in the teaching of mathematics is to help students develop \textit{mathematical competence}; that is the ability to understand, judge, do, and use mathematics across a variety of mathematical situations (\textit{Niss, 2007}). Basic mathematical competencies include problem-solving abilities (how to solve tasks without knowing a solution method in advance), reasoning ability (the ability to justify choices and conclusions), and conceptual understanding (insights regarding the origin, motivation, meaning, and use of mathematics). In an experimental design the present study primarily addresses whether and how students can develop conceptual understanding through mathematical problem solving and mathematical reasoning by engaging in more creative activities than procedure-based learning using predefined algorithms (e.g., \textit{Haavold, 2011; Lithner, 2003, 2008}). In addition, the mathematical task solving and reasoning are considered in relation to individual variation in cognitive proficiency. The present study is carried out in an experimental design and in that context it is important to point out that the proportion of studies that have been conducted pertaining to mathematics education, and that adopt experimental designs, is rare. During 2012, only 3\% of papers published in leading mathematics education journals used experimental designs (\textit{Alcock, Gilmore, & Inglis, 2013}).

\textsuperscript{*} Corresponding author. Tel.: +46 706777612; fax: +46 78666955.
E-mail addresses: bert.jonsson@psy.umu.se, bert.jonsson@live.se (B. Jonsson).

\url{http://dx.doi.org/10.1016/j.jmathb.2014.08.003}
0732-3123/\textcopyright\ 2014 The Authors. Published by Elsevier Inc. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/3.0/).
1.1. Learning in mathematics

Much time in mathematics classes is spent learning and rehearsing algorithms, which are supposed to provide students with a quick and reliable way to cope with many of the tasks ahead (Boesen et al., 2014; Hiebert, 2003). There are, however, doubts as to whether these algorithms actually give rise to any deeper understanding of the principles of mathematics, or whether the extensive use of algorithms is counterproductive (Hiebert, 2003). The notion of an algorithm includes all pre-specified procedures, that is, finite sequences of executable instructions that allow one to solve a given set of tasks (Brousseau, 1997). The importance of an algorithm is that it can be determined in advance, and the execution of an algorithm is associated with high reliability and speed, which is the strength of using an algorithm when the purpose of a task is only to produce an answer to a particular problem. In many cases, using an algorithm is appropriate; it saves time and prevents miscalculations. In this way, using algorithms provides students with opportunities to solve tasks simply by reusing the procedure that a particular algorithm stands for. However, the use of algorithmic reasoning is, in itself, not an indication of one’s conceptual understanding of mathematics (Haavold, 2011). In addition, the reason why an algorithm is regarded as efficient in solving a task (but not for learning) is that it is designed to avoid meaning (Brousseau, 1997). Algorithms are often presented within a classroom context. A typical situation arises whereby the teacher or textbook provides students with a set of mathematical tasks and a template solution method (algorithm); this is then followed by massive repetition of the algorithm, leading to an un-reflected use of the same algorithm (Boesen et al., 2014; Lithner, 2008). The tasks can, therefore, be solved according to the provided template without any conceptual understanding of the actual problem. Sufficient amounts of exposure to the algorithm may also lead to rote learning (the process of learning something by repeating it until it becomes memorized, rather than learning something by understanding the meaning of it; Oxford Advanced Learner’s Dictionary); this means that the algorithm can be recalled in its original form without any conceptual understanding of it.

In the present study, we define using memorized or well-rehearsed procedures (such as algorithms) without reflecting on their meaning as algorithmic learning. An important note is that using well-rehearsed procedures or engaging in rote learning can be an efficient way to learn facts such as multiplication tables (Caron, 2007). In a similar way, using algorithms can reduce the cognitive demands of complicated calculations (Haavold, 2011), and thus also the cognitive load on our working memory (Raghubar, Barnes, & Hecht, 2010).

The components and capacity of working memory refer to the ability to process and store information simultaneously (e.g., Baddeley, 2010). Students could, therefore, be aided by using algorithms that reduce the cognitive load, thereby freeing resources for more advanced problem solving to occur (Merriënboer & Sweller, 2005). However, if all or most learning is done using routine procedures, it can lead to algorithmic reasoning that is based on superficial features of the algorithm, and not on the intrinsic properties of the tasks at hand (Hiebert, 2003; Lithner, 2003); as a result, there is a risk that mathematical competences are not well developed. In spite of being efficient in the short term – in the sense that students can quickly solve new practice tasks, as long as there are templates to use and memorize – there are many studies showing that procedure-based teaching models fail to enhance students’ long-term development in basic mathematical competencies (see Hiebert, 2003 for an overview). Several other concepts are used in the literature to capture similar phenomena related to the dichotomy between superficial versus deep/true/conceptual mathematical learning. In the seminal book “Conceptual Knowledge and Procedural Knowledge” (Hiebert, 1986), Hiebert and Lefevre defined conceptual knowledge as a form of knowledge that is rich in informational relationships, and linked in a network where the connections within the network are as important as the discrete pieces of information themselves. Procedural knowledge was defined in terms of a person’s ability to become familiar with conventions of mathematics, while having access to the rules or procedures required to solve mathematical problems (Hiebert & Lefevre, 1986). However, Star (2005) argued that conceptual knowledge does not necessarily need to have a rich informational relationship. For example, a child’s conceptual knowledge can be less sophisticated and differently connected than that of an adult, but it is still regarded as conceptual knowledge. In a study by Rittle-Johnson and Alibali (1999) it was argued that ‘conceptual instructions’ (children were told the underlying principle behind the problem solution) to greater extent than procedural based instructions (being taught the procedure) influence conceptual understanding. However the results also indicated that the relationship is bidirectional, see also Rittle-Johnson, Siegler, and Alibali (2001) and Schneider, Rittle-Johnson, and Star (2011). In the present study, no ‘conceptual instructions’ such as the underlying principles are provided. The key issue in the present study is allowing for mathematical “struggle” in didactical situations (no teacher support) with tasks that are designed to facilitate students’ own construction of solutions.

1.2. The importance of a productive “struggle”

In order for students to obtain desirable learning outcomes, “the students need to be engaged in activities where they have to ‘struggle’ (in a productive sense of that word) with important mathematics” (Niss, 2007, p. 1304). At the same time, a delicate balance must be maintained in order to prevent these struggles from becoming obstacles, rather than promoters of learning. Hiebert and Grouws (2007) concluded in a mathematics education research review that this ‘struggle’ is necessary in order to enhance students’ development of conceptual understanding of the principles involved in mathematics. Still, little is known about how this idea of a ‘struggle’ translates into specific activities that are useful in the teaching of the subject, and in what way these activities are linked to learning outcomes (Niss, 2007). However, support for the argumentation of learning outcomes can be found in the field of memory research, where several studies have shown that more ‘struggle’ in terms of more effortful retrieval is effective for later performances on subsequent tasks (e.g., Pyc & Rawson, 2009); these are
results that have also been translated and proven effective in teaching (e.g., Carpenter, Pashler, & Cepeda, 2009; Karpicke & Blunt, 2011; Larsen, Butler, & Roediger, 2008; Wiklund-Hörnqvist, Jonsson, & Nyberg, 2014). Additional evidence comes from research on mnemonic strategies. Derwinger, Neely, and Bäckman (2005) showed that elderly that were encouraged to create and practice their own memory strategies, eight months later improved in a recall task for which there was no strategy support provided. Whereas the participants that during training where provided with a mnemonic strategy dropped in performance in the recall task eight month later.

1.3. Task design

If a task is appropriately designed it will, (a) promote students’ conceptual understanding of mathematics, while capturing and retaining their interest, and it will (b) optimize their learning (Chapman, 2013). An empirical example of how task design can be manipulated, and subsequently stimulate learning, is from the Kapur (2008, 2011) studies. In the context of productive failures (failures that occur during practice that also have positive effects on learning), participants had to practice on either ill-defined or well-defined tasks. An ill-defined task was a task that possessed unknown parameters, multiple solutions, and required the participants to make certain assumptions about the task. Well-defined tasks possess fewer parameters that can vary and, therefore, enhance participants’ confidence in completing the task (for a full explanation see, Kapur, 2008, 2011). For the ill-structured tasks, the challenge for the students is to extract the meaning of the task without any additional support. The results showed that participants who practiced on ill-structured tasks performed worse than those practicing on well-defined tasks. However, during the post-test, the pattern was reversed; participants who had practiced on ill-defined tasks outperformed participants that practiced on well-structured tasks across both well-structured and ill-structured tasks. It was argued that working with a higher degree of complexity and divergence (which could also lead to failures) facilitated the participants’ ability to develop structures that are helpful for problem solving; this was in contrast to well-structured problems for which those structures were already imposed in the task design. These results indicate that task design is important for enhancing mathematical reasoning, task solving and conceptual learning.

1.4. Task design and mathematical reasoning

Lithner (2008) suggested that a key variable in learning mathematics through task solving is the reasoning that students activate in relation to specific tasks. In the present study, two types of reasoning were addressed: algorithmic reasoning (AR) and creative mathematically founded reasoning (CMR). Lithner (2008) defined CMR as fulfilling all of the following criteria: (i) Creativity: a new reasoning sequence (new to the reasoner) is created, or a forgotten one is re-created, in a way that is sufficiently fluent and flexible enough to avoid restraining fixations; (ii) Plausibility: there are arguments supporting the strategy choice and/or strategy implementation explaining why the conclusions are true or plausible; and (iii) Anchoring: the arguments are anchored in the intrinsic mathematical properties of the components that are involved in the reasoning required to solve the problem. An important note is that the aspect of creativity that is emphasized in Lithner’s (2008) framework is neither ‘genius’ nor ‘exceptional novelty,’ but rather it is the creation of mathematical task solutions that are original to the individual who creates them, though the solutions can be modest. Although specific tasks or whole programs can require a high cognitive level (Stein & Kim, 2009), the mainstream teaching mainly promotes procedural-based learning (Bergqvist & Lithner, 2012; Boesen et al., 2014; Palm, Boesen, & Lithner, 2011). Judging from the research survey by Hiebert (2003), this may also be the case outside of Sweden, as observed (for example) in common American calculus textbooks (Lithner, 2004). In the present study, AR is defined as a repetitive numerical task-solving method that uses algorithmic support (i.e., an algorithm that can be used to solve the problem is provided together with the task). The opportunities for students to practice CMR are rare in teaching, textbooks, and tests, but when occasionally applied, the CMR approach has been found to be more efficient than AR for resolving certain problematic task solving situations (Lithner, 2008). However, one of the main problems is that students are seldom exposed to (and often do not need to use) CMR-based learning since the tasks can often be solved through AR. In addition, the effects from learning by CMR compared to AR have not been studied so far.

1.5. Didactical situations

In addition to how tasks are designed, it is important to create situations where the student (the learner) can construct the target knowledge (Brousseau, 1997), hence providing a situation allowing for “struggling with tasks solutions”. In this devolution of problems approach, students have to take responsibility for (a part of) the task-solving process. The teacher’s concern is to arrange a suitable didactical situation in the form of a problem. From the time when the student accepts the problem as his or her own, to the moment when he or she produces an answer, the teacher refrains from interfering. This part of the didactic situation is called an addidactical situation by Brousseau. The student must construct new knowledge and the teacher must therefore arrange for the devolution of a good problem, rather than communicating knowledge how to solve it. In this study, we designed tasks that constituted an addidactical situation during practice. The participants in the present study did not receive additional information or support other than information already present in each task respectively. The participants also worked by themselves, and they did not receive any help from their peers.
1.6. Individual variation in cognitive proficiency

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, such as the specific teaching situation and task design, but also individual variations in cognitive abilities and topic-specific knowledge. Working memory (WM) 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). 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 visuospatial 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 and as a measure of fluid intelligence which in turn is a prominent factor of general intelligence and an important predictor of mathematical achievement (Primi, Ferrão, & Almeida, 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 (e.g., Boesen, 2006). A third important factor in our study is topic specific knowledge (i.e., basic mathematical competence), which of course also is important for the mathematical performance. Altogether, it is evident that measures of WM, non-verbal problem solving ability and acquired topic-specific knowledge are central aspects that have a high impact on school achievements, and they can vary considerably between individuals. We thus measured both non-verbal problem solving ability and WM, and we collected grade nine students’ grades in mathematics at a final compulsory school level as a measure of topic-specific knowledge.

2. Aim and research questions

The present study addresses one of the most persistent problems in mathematics education: the replacement of dominating algorithmic-based teaching models with models emphasizing students’ own construction of knowledge. Learning is framed in an didactical teaching situation (Brousseau, 1997) promoting productive struggle using tasks that are designed to facilitate students’ own construction of solutions (Lithner, 2008). The purpose of the present study is to investigate the learning effects of practicing mathematical tasks through AR and CMR on task-solving performance while adopting the creative mathematical reasoning framework (Lithner, 2003, 2004, 2008) and an individual variations perspective of cognitive proficiency (Alloway et al., 2009; Alloway & Alloway, 2010; Conway, Kane, & Engle, 2003; Raghubar et al., 2010). Three hypotheses are tested:

(1) AR will lead to better performance when solving practice tasks, as compared to CMR, as a function of the algorithmic support provided (Lithner, 2008).

(2) CMR will outperform AR on time-limited tasks and on tasks that are aimed at (re)constructing task solutions (Lithner, 2008). The rationale is that practicing with CMR task emphasizing students’ own construction of solution in an didactical context (Brousseau, 1997) requires more effortful ‘struggle’ (Kapur, 2008; Niss, 2007) that results in better memory consolidation (e.g., Pyc & Rawson, 2009). In addition, the generalization that develops after self-generating the solution methods can facilitate conceptual understanding of the specific task solving methods and thus also enhance both memory retrieval and later (re)construction of a specific task solution.

(3) Cognitive proficiency will be significantly associated with performances. Participants with a higher cognitive proficiency will perform better (e.g., Adams & Hitch, 1997; Alloway, 2009; Andersson & Lyxell, 2007), irrespective of their study group (CMR versus AR).

3. Methods

3.1. Participants

For the study, 131 students agreed to participate. The participants were between 16 and 17 years of age. The students were recruited through contacts with headmasters and teachers in four upper secondary schools. Written informed consent was obtained in accordance with the Declaration of Helsinki, and the study was approved by the Regional Ethical Review Board, Sweden. The students completed the study within the school year, but outside of the ordinary curricula. Initially, 28 participants were excluded due to attrition, as they did not show up to practice or to the test session 1 week later. Additional four participants that performed below 10% during the practice sessions were assumed to not really have tried to solve the practice tasks and were therefore excluded. To prevent potential type I errors, participants scoring at ceiling were excluded (Austin & Brunner, 2003). The dependent variables were screened for outliers and, as a consequence, two more participants were excluded. Altogether, 91 participants were included in the analyses (48 in AR and 43 in CMR).
3.2. Materials

A set of novel training tasks was constructed to enhance students’ learning through AR or CMR. The tasks developed for the present study are based on extensions of a previously developed method (Boesen, Lithner, & Palm, 2010; Palm et al., 2011) that, through analyses of the mathematics textbooks that were used, estimates a student’s (or a student group’s) likelihood of using CMR or AR. In the following sections, the method is outlined and the task design is described.

3.3. Practice tasks

The target knowledge for both training groups was solution methods for 14 different mathematical task sets (Fig. 1 exemplifies 2 of the 14 task sets). ‘Solution method’ refers to a method to solve a particular task that is applicable to different numerical input values. For example, if the task is to find out how many matches are needed to form a row of squares
(Fig. 1b and d) then one solution method is ‘form a mental image of the squares and count the matches’, a second is ‘start by one match and add three new for each square’ and a third is ‘the number of matches is 3x + 1 where x is the number of squares’.

**AR:** Students in the AR group were given five numerical sub-tasks for each of the 14 task sets. For AR tasks the solution methods for each task set was provided in the form of an algebraic formula, as well as an example on how to apply it (Fig. 1a and b shows examples of two different AR tasks sets). Hence, for each of the 14 task sets, the participants were provided with a correct formula for all the five consecutive sub-tasks.

**CMR:** Students in the CMR group were given three sub-tasks for each of the 14 task sets, with no guidance provided on how to solve the tasks (Fig. 1c and d shows examples of two different CMR tasks sets). During practice the participants had to (i) create a new reasoning sequence in each task set, and since it is unlikely that the tasks can be solved by pure guesswork the participants had to (ii) reflect on whether their solutions are true, or at least plausible, and whether the reasoning is (iii) anchored in intrinsic mathematical properties of the components that were involved in the reasoning (c.f. i–iii in the CMR definition above). In the third sub-task the participants were asked to generate a mathematical formula (function) based on the previous two sub-tasks (Fig. 1e and f). As per its definition, self-generation produces a generalized knowledge (a basic principle) of the numerical task being practiced.

The AR approach is similar to what is usually provided through teaching and textbooks. AR is thus regarded as a teaching baseline against which CMR is compared. In an attempt to equate practice time, there were more AR practice sub-tasks than CMR practice sub-tasks (five and three, respectively). See Fig. 1a–f for examples of practice tasks.

### 3.4. Test tasks

The test tasks were identical for both the AR and CMR groups, and the sub-tasks for each of the 14 task sets were denoted as test tasks I–III. In test task I (formula), the participants were asked to write down the formula corresponding to the practice task; the time limit was set to 30 s. Test task II (short numerical) was comprised of numerical tasks with the same time limit as in test task I. The idea is that there was enough time to recall and apply a solution (e.g., a principle for finding the number of matches), but not enough time to re-construct the task. In test task III (long numerical), the same numerical task from test task II was presented, but now with a 300-s time limit. This allowed for (re)construction of the solution. The test tasks II and III were identical to the CMR practice tasks 1 and 2 (Fig. 1c and d), but with different numbers. Concerning the time limit for tasks I and II pilot studies indicated that when a student in advance knew how to solve this type of task, 30 s was sufficient. In addition, estimates indicated that it took some 20 s just to read the task and write the answer. The remaining 10 s was judged to be sufficient to recall an answer or solution method, but not to (re)construct it. Furthermore, pilot studies indicated that in most cases were students managed to construct new solution methods (without time limits), 300 s was sufficient. Thus the idea is that test I tests if the student can recall a solution formula, test II tests if any solution method can be recalled and applied and test III tests if any solution method can be (re)constructed. One may note that if test task I was solved correctly, then the same formula could be used to solve test tasks II and III. However, test task III is also possible to solve through a (re)construction (in the same way as in the CMR practice session) without any need to recollect a formula or solution method.

### 3.5. Cognitive measures

Participants were initially measured on two cognitive variables: Raven’s Advanced Progressive Matrices (APM) (Raven & Raven, 1991) and Operation Span (Unsworth, Heitz, Schrock, & Engle, 2005). These are measures of non-verbal reasoning and working memory capacity, respectively. Operation Span has good test–retest reliability and internal consistency (e.g., Conway, Cowan, Bunting, Therriault, & Minkoff, 2002; Engle, Tuholski, Laughlin, & Conway, 1999; Klein & Fiss, 1999), and it correlates well with other measures of WMC and higher-order tasks (Conway et al., 2002; Unsworth & Engle, 2005). Operations Span is computer based and the responses are entered using a keyboard. Participants are required to solve a simple mathematical task while trying to remember a letter presented immediately following the operation. Immediately after the letter is displayed, a new mathematical task is presented. The operation and letter are presented in a set of two to seven items. Following each complete set, the participants are required to recall the letters in the order that they were presented. In the present study, the number of accurately recalled letters was scored.

In Raven’s APM 36 items are presented in ascending order, and each item consists of a 3 × 3 matrix of geometric patterns with the bottom-right area missing a pattern. The task is to complete the pattern by selecting one option among eight alternatives. In the present study, 18 items presented in ascending order of difficulty were used. Participants were allotted 20 minutes to complete the task, and the numbers of correct solutions were scored (to a maximum of 18). The order of presentation for Raven’s APM and for the Operation Span was outbalanced among the participants within each group. The correlation between Raven’s APM and Operation Span was found to be 0.43, \( p < .001 \). Raven’s matrices and Operation Span were, therefore, transformed to z-scores and used together as a composite score of cognitive proficiency, a Cognitive Proficiency Index (CPI).
Table 1
The intercorrelations between practice tasks for AR (1–5) and for CMR (1–3), respectively.

<table>
<thead>
<tr>
<th>AR</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>CMR</th>
<th>1</th>
<th>2</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Practice task 1</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Practice task 2</td>
<td>.62**</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td>.70**</td>
<td>-</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Practice task 3</td>
<td>.62**</td>
<td>.71**</td>
<td>-</td>
<td></td>
<td></td>
<td>.67**</td>
<td>.85**</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>Practice task 4</td>
<td>.57**</td>
<td>.70**</td>
<td>.64**</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Practice task 5</td>
<td>.50**</td>
<td>.71**</td>
<td>.81**</td>
<td>.56**</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

** p < .01.

3.6. Design

In a mixed factorial design, the group (AR or CMR) was manipulated between subjects, whereas performances on the practice and test tasks were manipulated within subjects. Background variables such as age and sex were collected together with the final grade in mathematics from the last year in compulsory school. CPI was together with mathematical grade and sex used to match participants into two separate groups. Thus, the groups were considered approximately equal in terms of cognitive proficiency, mathematical prerequisites, and sex distribution. Each group was assigned to a set of training tasks that led to either AR or CMR.

3.7. Procedure

Measures of cognitive proficiency were collected one week before the practice session, allowing time for matching participants into two groups. Both the practice and test sessions were run in a standard web browser. Performance data were automatically stored by the same software that presented the practice and test tasks. The participants were trained on either the AR or CMR tasks in one session. After 1 week, the participants returned for tests. The allowed time for both the practice and test sessions was equal for both groups (210 and 85 min, respectively). The amount of practice time that was effectively used by the participants was, on average, 29 min (SD, 10) for CMR and 21 min (SD, 6) for AR. All participants worked individually during the practice session, and they were subsequently tested individually. An experimenter was present in the lab during both the practice and test sessions to monitor the procedure. During the practice session, no information was given about the content of the upcoming sessions. No assistance was provided, except for answers to questions about how to use the computer. This was done to ensure that the structure of the experiment reflected an adidactical situation; in this way, the student must construct new knowledge about the task without influence from a teacher (Brousseau, 1997). The dependent measures of practice and test tasks performances are all based on the proportion of correct responses. A composite score included the mean value of the 14 task sets that were used to arrive at a solution during the practice or test sessions, and this score was ultimately derived from each sub-task. For the test tasks we also calculated the average score for each sub-task (I, II and III).

3.8. Data screening

The students had three levels of grades from school (pass, pass with distinction and past with honor), however there were only five students with the grade level pass (one AR and four were CMR) and those students were therefore amalgamated with the next level (pass with distinction). There were 25 students with a first level grade (pass with distinction) and 23 with a second level grades (pass with honor). Corresponding values for CMR were 19 and 24, respectively. The Shapiro–Wilks test of normality showed that CPI scores violated the assumption of normality and was therefore transformed using a squared transformation. After the transformation the variable was normally distributed, S–W = 0.98, df = 90, p = .47. The composite practice scores was also found to be non-normally distributed but could not be corrected through transformation. Composite practice score was therefore excluded in the parametric analyses. However, to investigate the group difference in practice task performances a non-parametric test was conducted. Additional screening revealed a skewness (>1) and kurtosis (above three standard errors) for the amount of practice and was therefore log_{10} transformed. After the transformation the skewness and kurtosis were 0.43 and 0.39 respectively with corresponding standard errors of 0.25 and 0.50.

As was the case in the initial sample, there were more females than males in the remaining sample, but there was an equal distribution across both groups (AR, 16 males, 32 females, CMR, 13 males, 30 females). T-tests showed that there were no sex-differences across any of the independent or dependent measures, all p > .31. Sex was therefore excluded from further analysis.

For both AR and CMR, the practice tasks were significantly and highly correlated with each other (see Table 1). Composite scores based on the mean values of the AR and CMR practice tasks were therefore formed and are denoted “composite practice AR” and “composite practice CMR.” Similarly, the test tasks (formula, short numerical, long numerical) were highly correlated and significant (Table 2). The composite scores for the tree test tasks were denoted as “composite test AR” and
Table 2

The correlations between test tasks (I–III) for AR and CMR respectively.

<table>
<thead>
<tr>
<th>Test tasks</th>
<th>AR</th>
<th>CMR</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>Test task 1</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>Test task 2</td>
<td>.91**</td>
<td>–</td>
</tr>
<tr>
<td>Test task 3</td>
<td>.76**</td>
<td>.86**</td>
</tr>
</tbody>
</table>

** p < .01.

"composite test CMR." However, levels of performance across both the CMR and AR test tasks (I–III) differed; consequently, the test tasks were also analyzed separately.

3.9. Statistical analyses

Practice task performance (composite scores) for AR and CMR was compared using a non-parametric test (Mann–Whitney) and test task performance (composite scores) was evaluated using an independent t-test. To pursue the question of whether practice scores are related to test task performances and group we conducted a non-parametric partial Spearman rank correlation analysis (Conover, 1980) between the composite practice score and composite test tasks scores when controlling for group. Significant effects were followed by Spearman rank correlation analyses for AR and CMR groups separately. To evaluate the effect of practice on the three separate test tasks we conducted a 2 × 3 mixed-model analysis of variance (ANOVA) with group (AR versus CMR) as the between-subject factor and with the test task (I–III) as the within-subject factor. Homogeneity of variance was checked to ensure no violation of the assumption. Greenhouse–Geisser was used to correct the degrees of freedom in case Mauchly's test indicated that the assumption of sphericity was violated. The corrected degrees of freedom were rounded up to nearest integer. To investigate the independent variables as predictors of performances we entered the amount of practice and CPI in a regression analysis. To examine potential moderating effects the interaction terms Group × Amount of practice, Group × CPI and Group × Grades were also entered as predictors in the regression analysis. Before running the regression analysis we followed the suggestion by Aiken and West (1991) that continuous variables included in an interaction term should be mean centered in order to decrease collinearity. A subsequent collinear diagnostic of the mean centered predictor variables revealed no Variance Inflation Factor (VIF) value above 3, hence no risk for multicollinearity. All statistical analyses were conducted using the Statistical Package for the Social Sciences, version 22 (SPSS 22) except for the partial rank correlation for which Conover (1980) formula was used and calculated in Excel.

4. Results

4.1. Practice and test tasks performances

Fig. 2 shows the composite practice and test scores and the three test tasks (which were denoted as formula, short numerical, and long numerical). The composite scores in Fig. 2a clearly shows that AR outperformed CMR during the practice session and that CMR group outperformed the AR group during the test. A Mann–Whitney test comparing the composite practice scores between the CMR and AR groups confirmed that those in the AR group outperformed those in the CMR group, $U = 106.5$, $p < .0001$, $r = .77$. An independent samples t-testing comparing the composite test scores between the CMR and AR showed that the pattern was reversed: CMR significantly outperformed AR, $t(89) = 3.54$, $p = .001$, $d = 0.73$. Fig. 1b shows the test scores for formula, short numerical and long numerical tasks. The mixed-model ANOVA with the test tasks (I–III) as a within-subjects factor, and group (AR versus CMR) as a between-subjects factor revealed a main effect of test task, $F(2, 178) = 139$, $p < .0001$, $η^2 = .61$, and group, $F(1, 89) = 12.52$, $p < .001$, $η^2 = .12$. These main effects was not qualified by an interaction between test task and group, $F(2, 178) = 0.60$, $p = .55$, $η^2 = .007$, showing that the group difference was stable across the three test tasks. Fig. 2 shows the average practice and test task performances (proportion of correct responses) for the three tasks (short formula, short numerical, and long numerical) for both the AR and CMR groups.

4.2. The association between practice and test task performances

To analyze the association between composite practice scores and composite test tasks scores when controlling for group, a Spearman rank Correlation analysis was conducted. The results revealed a significant effect $r_{xy:a} = .46, n = 89, p < .001$. Indicating that the correlations differ with respect to group. For that reason separate Spearman rank correlation analyses were conducted. For AR the correlation was non-significant, $r_{xy} = .25, n = 48, p = .10$, for CMR however the correlation was highly significant, $r_{xy} = .81, n = 43, p < .0001$. The analyses indicate that more correct responses during practice is associated with test task performances and that this association is stronger for CMR participants.
4.3. Regression analysis

To estimate the impact of each independent variable in relation to all of the other variables, we entered group, CPI, Amount of practice as main effect predictors and Group × CPI and Group × Amount of practice as potentially mediators. The predictors were entered simultaneously and the composite test scores were entered as the dependent variable. The regression analyses (Table 3) showed that Group, CPI and CPI × Group were significant predictors. The results also showed that the seven predictors explained 43% of the variance. Table 3 shows the explained variance (F and p values) for each test task with predictor-specific beta values, standardized beta values (t-test values), and corresponding levels of significances. Fig. 3 shows the relationship between performance and CPI for each group. The figure shows that there is a main effect of CPI, but also that the regression slope is steeper for AR participants than for CMR, i.e. the interaction effect.

5. Discussion

In the present study, we hypothesized that AR would lead to better performance during the practice session (hypothesis 1) when compared to CMR, as a function of the algorithmic support that was provided. The results showed that AR indeed outperformed CMR during practice.

In hypothesis 2, it was argued that CMR would outperform AR. The argument were that CMR requires more effortful processing, and that the generation of a solution method leads to a generalization of the tasks at hand, thereby facilitating conceptual understanding, memory retrieval and/or (re)construction of solution methods. The results showed that CMR outperformed AR in all three test tasks. In hypothesis 3, it was argued that CPI would be significantly associated to test task performance. The hypothesis was confirmed and is in line with those from many studies that have investigated the relationship between cognition and school achievements (e.g., Alloway & Alloway, 2010; Andersson, 2008; Ashcraft & Krause, 2007). However the regression analysis also showed that the effects were more pronounced for the AR group (Fig. 3).
The results are elaborated upon and discussed in terms of practice and test task performances, mathematical struggle, generalization of knowledge, memory retrieval, cognitive proficiency, and didactical situations. In addition alternative explanations are discussed.

5.1. Practice and test task performance

The initial analyses of composite practice scores showed, as expected, that AR outperformed CMR, arguable from the support of the provided formulas. From Fig. 2a and b, and in the corresponding analyses, it is clear that the pattern of performance during the practice session and the test tasks was reversed. However the reverses pattern was not analyzed in terms of within subject decline. The practice sessions are as pointed out manipulated to be different in algorithmic support or not, and there was in the present study no pre-measure that are equal for both groups that could be uses in a clear cut within-subject evaluation of the decline from practice to test. In addition, the practice test performances were found to violate the assumption of normality and was therefore excluded from the parametric analyses. To pursue the question of the association between practice and test tasks we therefore conducted non-parametric correlation analyses (Spearman rank correlation) between practice and test tasks. For CMR but not for AR the association between frequency correct responses and later test performances was found to be highly significant, indicating that frequency correct responses during practice was to a greater extent associated with test performances one week later for CMR.

5.2. Test task I–III performances

In test task I, the participants had to recall a formula, since it is unlikely that they could read the task, re-construct the solution method, and write down the solution in 30 s. CMR was found to outperform AR. For test tasks II and III (a numerical task completed within 30 and 300 s, respectively), CMR still outperformed AR, but the proportion of correct responses improved equally for both groups. This results and the high correlation between tests tasks (I–III) indicate that the test tasks to a high extent measure the same underlying phenomenon. Further studies will disentangle the relation between types of tasks. It is important to note that if test task III performances were driven by construction from scratch, the AR performances should approach similar levels to CMR performances, considering the extensive time limit of 5 min. This is, however, not the case. This indicates that higher performance among the CMR group on test task III when compared to AR was at least partly driven by a higher degree of re-construction, and it was likely facilitated by a conceptual understanding that was consolidated during practice and reinstated at test task I and II.

5.3. Mathematical struggle and generalization of knowledge

As pointed out above, it is clear that CMR participants performed at a much lower level during practice and that they also spent more time in practicing, hence struggling with the tasks. This could indicate that engaging more effortfully is a key for later performances. Those arguments are in line with the effortful retrieval hypothesis (Bjork, 1994; Pyc & Rawson, 2009; van den Broek, Takashima, Segers, Fernández, & Verhoeven, 2013), arguing that more effortful encoding facilitates later performance and with Niss’ (2007) argument that struggling with important mathematics is necessary for subsequent performance. Although CMR participants allocated significantly more time for practice, it is important to note that the amount of practice” or “Group × Amount of practice” did not emerge as significant predictors in the regression analysis. The
Spearman rank correlation analyses indicate that it is not “time on task” that explains the struggling; rather, it is whether the amount of struggle faced with during practice leads to a correct answer or not at test. In addition there are studies showing that also failure during practice could be beneficial for later performance (e.g., Kapur, 2008; Richland, Kornell, & Kao, 2009), this aspect was however not investigated in the present study.

A potential explanation to why CMR outperform AR is that the generation of the formula (CMR practice, sub-task three) consolidated the formula as a specific long-term memory. An alternative explanation, although not mutually exclusive, is that the generation of a specific formula created a generalization (a basic principle) of the formula. The latter hypothesis is in line with Dalhberg and Housman’s (1997) findings that generating examples is effective for attaining an understanding of new concepts. The present design did not allow for the disentangling of formula generation from the numerical tasks themselves.

5.4. Predictors of performances

In order to investigate each independent variable in relation to all of the other independent variables as predictors of performances they were all entered in a linear regression analyses. Since there was no strong theoretical assumption about the order of entering the predictors, they were all entered simultaneously. Group, CPI and Group × CPI, were found to be significant predictors for the composite test scores. The results are in line with the results from the ANOVA and with the assumptions made about an adidactical situation (Brousseau, 1997) and Lithner’s (2008) model of creative mathematically founded reasoning. This is in contrast to an AR setting, in which students can rely on the algorithms themselves, thus preventing their development of a conceptual understanding of mathematical principles (Lithner, 2008). The main effect of cognitive proficiency was expected since the tasks are cognitively demanding. However the effect of the Group × CPI interaction (see Fig. 3) shows that the cognitive demands are significantly higher for AR participants when being tested 1 week later. It should be stressed though, that the sample in this study was rather homogenous (all participants were from natural science programs), and may not be representive for more marked interindividual cognitive proficiency variability in more heterogeneous samples. Additional studies of more heterogeneous groups will be needed to further evaluate the significance of cognitive proficiency.

5.5. Alternative explanations

An alternative interpretation that might partly explain the present results is the transfer-appropriate processing (TAP) view (Morris, Bransford, & Franks, 1977; Tulving & Thomson, 1973 c.f., the encoding-specificity principle). The TAP perspective states that if there is a close relationship between how information is initially encoded and subsequently retrieved, performance on tests is facilitated. In the present study, the practice tasks for AR and test tasks II and III only differed in terms of the formula always being displayed during the practice tasks and not during the test tasks. For the CMR group, practice tasks I and II, and test tasks II and III were identical, except for the fact that the numbers used in each sub-task differed. If a person deliberately and intentionally had to construct their knowledge, as in the CMR practice tasks, it is not unlikely that recalling a solution method (test task II) and later construction of re-construction (test task III) is facilitated as a function of similar underlying processes. However, participants were, in addition to the formula, exposed to exactly the same information as those in the CMR group; therefore, it seems as TAP is an unlikely explanation for the differences in CMR/AR performances. It could further be argued that test task I might use the same underlying process that are required when generating the formula during CMR practice task III. When being asked to retrieve the formula (test task I), it is possible that the memory traces – which were established when the formula was generated are reinstated. Nevertheless, as Karpicke and Zaromb (2010) pointed out, self-generation cannot be regarded as involving the same processes that were used when explicitly being asked to retrieve memory-based information.

5.6. Limitations

There are some limitations in the present study. There was no control over the participants’ activities in the week between the practice and test sessions. Still, it seems unlikely that students practiced similar tasks at school or at home. They were also instructed to not discuss the tasks with each other. Both the practice and test sessions were conducted outside of the normal curriculum, and it is therefore difficult to estimate the implications of this, as it compares to task solving within regular class time. However, the circumstances were the same for both groups. Although we tried to mimic classroom settings using educational relevant tasks it is however worth remembering that the study was conducted in an experimental setting and that the ecological validity could therefore be questioned.

6. Summary

In summary, it is shown in the present study that the CMR approach was more effective than the AR approach in terms of memory retrieval and construction of knowledge. The study also showed that the effect of cognitive proficiency was more pronounced at test for participants previously practicing in an AR setting, indicating that CMR is more beneficial for students cognitively less proficient. The main message is that in order for students to gain conceptual knowledge, it is important that
they are placed in an adidactical situation that provides them with opportunities to struggle with important mathematics (Brousseau, 1997; Niss, 2007). It is also important to give students the opportunities to create their own solutions to problems (Lithner, 2008). These results are in contrast to the common belief that cognitively less proficient students should not be involved in problem solving; that is, in order to overcome their limitations, they should learn algorithmic rules by rote learning instead (Boesen, 2006). To put it bluntly, all students should be given the opportunities (and perhaps be “forced”) to struggle with certain tasks. However, it is also worth revisiting the statement from the introduction, in that there is a delicate balance that must be achieved in order to prevent these “forced” struggles to become obstacles, rather than promoters, of learning.

Acknowledgments

This research was supported by Umeå University. We thank Tony Quilllard for help with parts of the data collection and computer programming.

References


