Learning mathematics without a suggested solution method: Durable effects on performance and brain activity

Article *in* Trends in Neuroscience and Education · April 2015

DOI: 10.1016/j.tine.2015.03.002

6 authors, including:

Bert Jonsson
Umeå University

Yvonne Liljekvist
Karlstads Universitet

Mathias Norqvist
Umeå University

Trends in Neuroscience and Education 4 (2015) 6–14

Contents lists available at ScienceDirect

Trends in Neuroscience and Education

journal homepage: www.elsevier.com/locate/tine

Research Article

Learning mathematics without a suggested solution method: Durable effects on performance and brain activity

Linnea Karlsson Wirebring [a,b,e,*], Johan Lithner [c,d], Bert Jonsson [e], Yvonne Liljekvist [f,g], Mathias Norqvist [d,h], Lars Nyberg [a,b,i]

[a] Department of Integrative Medical Biology, Umeå University, 901 87 Umeå, Sweden
[b] Umeå Center for Functional Brain Imaging, Umeå University, 901 87 Umeå, Sweden
[c] Department of Science and Mathematics Education, Umeå University, 901 87 Umeå, Sweden
[d] Umeå Mathematics Education Research Centre, Umeå University, 901 87 Umeå, Sweden
[e] Department of Psychology, Umeå University, 901 87 Umeå, Sweden
[f] Department of Mathematics and Computer Science, Karlstad University, 651 88 Karlstad, Sweden
[g] The Centre of Science, Mathematics and Engineering Education Research, Karlstad University, 651 88 Karlstad, Sweden
[h] Department of Mathematics and Mathematical Statistics, Umeå University, 901 87 Umeå, Sweden
[i] Department of Radiation Sciences (Diagnostic Radiology), Umeå University, 901 87 Umeå, Sweden

ARTICLE INFO

Article history:
Received 5 September 2014
Received in revised form
20 February 2015
Accepted 9 March 2015

Keywords:
Mathematics
Learning
fMRI
Parietal cortex
Angular gyrus
Education

ABSTRACT

A dominant mathematics teaching method is to present a solution method and let pupils repeatedly practice it. An alternative method is to let pupils create a solution method themselves. The current study compared these two approaches in terms of lasting effects on performance and brain activity. Seventy-three participants practiced mathematics according to one of the two approaches. One week later, participants underwent fMRI while being tested on the practice tasks. Participants who had created the solution method themselves performed better at the test questions. In both conditions, participants engaged a fronto-parietal network more when solving test questions compared to a baseline task. Importantly, participants who had created the solution method themselves showed relatively lower brain activity in angular gyrus, possibly reflecting reduced demands on verbal memory. These results indicate that there might be advantages to creating the solution method oneself, and thus have implications for the design of teaching methods.

1. Introduction

One of the fundamental cognitive skills an individual has to learn to master during development is the ability to reason logically with numbers. In fact, the ability associated with mathematical understanding during school age has been found to be highly predictive of success later in life (e.g. [1]) while poor mathematical skills can have negative consequences for the individual (e.g. [2,3]). Not surprisingly, mathematics is prioritized as a core subject in all school systems, from kindergarten to college, and countries' educational qualities are consistently evaluated and compared not least on the basis of pupils' mathematical performance (e.g. TIMSS and PISA international surveys). Recently, the neurosciences have witnessed an increase in the number of studies targeting learning of arithmetics (for one review, see e.g. [4]).

How can an educational system assure that mathematics is being taught in a way that most efficiently promotes mathematical learning? This is an area of extensive debate [5–7]. What has been observed in detailed analyses of mathematics textbooks and curriculums is that one dominant mathematical teaching method centers on presenting typical task types and then give suggestions for solution methods (for examples from Sweden, see: [8,9], and from the US: [7]). These suggestions for solutions, commonly in the form of algorithmic templates (e.g. rules, methods, solved example tasks: [10]), are then typically subjected to extensive repeated practice, for example via practice tasks throughout a book chapter. A typical introductory example task in a chapter on percentages could read: “Of 80 students finishing grade nine, 16 applied for the natural science upper secondary program. How many percent of the students was that?” This is then followed by a template solution and the correct answer: “Proportion of applicants: 16/80 = 0.20 = 20% Answer: 20% of the students applied for

* Corresponding author at: Department of Psychology, Umeå University, #901 87 Umeå, Sweden. Tel.: +46 70 567 95 91/+46 90 786 74 08.
E-mail address: linnea.karlsson@umu.se (L. Karlsson Wirebring).

http://dx.doi.org/10.1016/j.tine.2015.03.002
2211-9493/

the natural science program." Finally, this is usually followed by many practice tasks that are isomorphs to the introductory task, for example: "At a traffic control outside a school it was found that 84 cars out of 400 drove too fast. How many percent was that?"

Such teaching methods are guaranteed to lead to learning in the short term but conceptually, they appear to have much in common with 'rote learning': the process of learning something by repeating it until you remember it rather than by understanding the meaning of it (cf. Oxford Advanced Learners Dictionary). However, in spite of being short-term efficient there are data indicating that teaching based only on such methods fail to enhance students' long-term development of conceptual understanding [7]. Throughout this paper, we will refer to mathematical teaching methods of this kind as methods inviting Algorithmic reasoning (AR) [10].

As an alternative, it has been suggested that encouraging the individuals to create a solution method themselves should be superior for promoting mathematical learning, compared to explicitly presenting the solution method and invite extensive repeated practice with it [5]. This suggestion has been further specified by Lithner and colleagues by designing practice tasks inviting Creative Mathematically founded Reasoning (CMR) [8,10,11]. To compare with the example above, a task inviting CMR would include the same type of task, for example: "At a traffic control outside a school it was found that 84 cars out of 400 drove too fast. How many percent was that?"—but would not be preceded by the solved introductory task and template solution. Moreover, instead of being followed by many practice tasks, a task inviting CMR would instead be followed by explicit encouragement to create a solution method (e.g. a formula for the solution of the task). Jonsson et al. [11] have recently demonstrated that practice tasks designed to invite CMR might have superior effects on performance compared to tasks designed to invite AR.

The purpose of this study was to further compare these two kinds of practice tasks—designed to invite AR and CMR, respectively, both in terms of performance as well as in terms of brain activity. Participants first trained in an environment where they solved numerical tasks *with* given solution methods (AR), or solved numerical tasks *without* given solution methods (CMR). One week later they were tested on similar numerical tasks without given solution methods while being scanned with functional magnetic resonance imaging (fMRI), which allowed comparing the two teaching methods in terms of their effects on mathematical performance as well as on brain activity. A central question was whether these two kinds of practice tasks give rise to performance differences in the long-term [11].

Key to an environment designed to invite CMR rather than AR is that the solution method is not given but has to be self-generated [5,10]. Cognitive psychology research show that generating an answer compared to just reading it has large positive effects on long-term retention of that material, an effect known as the *generation effect* (e.g. [12]; see [13] for a review). This effect is related to the *testing effect*: repeated testing on a content for learning has stronger effects on long-term retention compared to repeatedly studying the same content (e.g. [14]). The generation effect has also been demonstrated with mental arithmetic [15,16]. For example, it has been shown that more answers to multiplication problems are remembered after previous practice on generating the answers compared to just reading the problem together with the answer [15]. Further, the benefit of generating the arithmetic solution has been shown to be larger for participants with low prior knowledge [16]. Even though the generation effect and the testing effect are empirical phenomena, with a wide range of potential theoretical explanations (see e.g., [13,17–20]), the demonstrated long-term performance benefits after self-generation are robust and compelling.

If participants trained in the CMR environment will have an easier time accessing their knowledge of a solution method during a later test, are there reasons to believe that this is manifested in relative differences in brain activity? To date, imaging studies of mathematics have in part focused on arithmetic tasks such as one- or two-digit multiplication, subtraction or addition tasks [4] or on task solving with algebra (see e.g., [21]) also for advanced algebra (e.g. [22]). Less is known about to what extent relative differences in brain activity observed in such tasks are also evident during less constrained and more general solution modes, as in, for example, creative mathematically founded reasoning.

One central aspect of practice effects in mental arithmetic that has gained much attention in imaging research is the shift as a function of practice from procedural calculation operations to retrieval of stored facts from memory (cf. [23]). Combining neuroimaging with multiplication tasks, for example, this shift has been observed to be mirrored by a relative shift in activity from frontal areas to parietal areas, in particular to the *left angular gyrus* (see e.g., [24,25]; for a review see [4]). Angular gyrus plays a key role in a model for number processing [26] and has been shown to be important for operations that in general require access to verbal memory of arithmetic facts, potentially supporting the verbal aspects of mental arithmetic. Thus, to the extent that participants trained in the CMR environment will require less effort to retrieve their knowledge of a solution method at the one-week follow-up test, we expect relatively *lower* activity in left angular gyrus in the CMR compared to the AR condition.

Finally, in this study we also investigated the potential role of individual differences in cognitive abilities of relevance for mathematical performance. It has previously been suggested that working memory is one potent predictor of mathematical achievement (e.g. [27]). Can individual differences in working memory explain variance in mathematical performance over and above potential effects of the different practice tasks? It has also been demonstrated that another potent predictor of mathematical competence is the acuity of the *Approximate Number System* (ANS: see e.g. [28,29]). The ANS is said to represent numerical magnitude in a non-symbolic mode. In order to investigate which cognitive abilities – if any – that explain variance in performance over and above the practice tasks, we included measures of *working memory* (Operation Span) and *ANS acuity*, as well as measures of *vocabulary* (SRB1) and *visuo-spatial processing and integration* (Raven's advanced matrices).

We hypothesized that ? (a) *creative mathematically founded reasoning* (CMR) will promote *better* performance at test one week after training than *algorithmic reasoning* (AR) [11], (b) CMR will translate into less engagement of left angular gyrus at test compared with AR, and (c) that working memory and ANS acuity will be significant predictors of individual differences in mathematical performance (independent of teaching method).

2. Methods

2.1. Participants

Seventy-three pupils and students participated in the study. The pupils (n=40, 24 males, 18–20 years old, M_{age}=18.5 years, SD_{age}=0.60) studied in their last year of the Swedish "gymnasium" (compares with upper secondary school/senior high school) and were enrolled in programs with a focus on natural sciences. The students (n=33, 24 males, 18–22 years old, M_{age}=20.0 years, SD=1.2) were enrolled in different engineering programs, in their first semester, and had all completed advanced mathematics courses during the gymnasium. All participants were right-handed and had normal or corrected-to-normal vision. Participants signed a written informed consent before participation. The study was approved by the Umeå University local ethics committee.

2.2. Design, procedure and materials

Participation included one session of individual difference measures, one mathematical training session and one mathematical test session. The mathematical training was done in one of two environments as a between-group design: participants either trained in one environment that was designed to invite algorithmic reasoning (AR) or in one that was designed to invite creative mathematical reasoning (CMR). The test session was done in the fMRI scanner. In order to reveal the possible lasting effects of mathematical training on performance, the test session was administered 6 days after the training session.

2.2.1. Individual difference measures

About one to three weeks prior to the mathematical training participants completed a battery of individual difference measures. These included a short form (18 items) of the *Raven's advanced matrices* [30], an automated *operation span task* (OSPAN; [31]), a *vocabulary test* (SRB1, [32]) and a test of the *approximate number system* (ANS; [28]). Moreover, participants completed a form with demographic variables, including information about their latest mathematics grade in upper secondary school/senior high school. Afterwards, participants were matched to two experimental conditions (AR and CMR), conditional on the groups being matched on Raven's score, latest mathematics grade and gender. About half of the pupils were allotted to the AR condition ($n=19$; 12 male) and the other half ($n=21$; 12 male) to the CMR condition. Half of the students were allotted to the AR condition ($n=17$; 12 male) and $n=16$ (12 male) to the CMR condition.

2.2.2. Mathematical training

Participants completed practice instances of nine mathematical task sets (see Table 1). The task sets were carefully tested in a separate study to ensure that they induced reasonable levels of performance irrespective of condition (see [11] for details). The nine task sets were in two different versions, one version designed to invite AR and one version designed to invite CMR. The tasks asked for a numerical solution to problems where it would be helpful to have a solution method at hand, as the constituent variables in each practice instance of a task set grew too large for feasible simple mental arithmetic. As an example, one of the task sets involved computing how many matchsticks would be needed in order to create one row of x squares of matches, where x varied between 6 and 100 in the practice instances.

In both versions of the task sets, the task was first introduced in general terms, together with an illustrative figure and a concrete question (see [11] for an illustration of the task). With the matchstick task for example, it was stated that: "When you put together squares in a row it looks like in the figure. For four squares in a row one will need 13 matchsticks." The question in the first instance was: "How many matchsticks do you need in order to put together one row of 6 squares of matchsticks?" Participants were carefully instructed to try to solve every instance and to press ENTER when they had figured out the correct answer. In cases when they were sure that they would not be able to solve it, they were instructed to press the SPACE bar to move on the next instance.

After pressing ENTER participants were given multiple response alternatives in an alternative forced-choice task. They were shown three numerical response alternatives and a fourth alternative "None of the above" (which was never correct). Participants marked which of the alternatives was the correct one and received feedback. After each response participants were asked to rate how confident they were that they were going to solve the instance when they first read it, on a scale from 1 ("not at all sure I would make it") to 5 ("totally sure I would make it"). Participants encountered a few instances of each task set before moving on to the next, with the value of x in the question differing from instance to instance.

A few key design features administered to this basic task setup were done in order to form a typical AR or CMR task, respectively.

The version of the task designed to invite AR had additional information below the general formulation of each task. A solution method adequate to solve the particular task set was presented (and in the first instance of the task set together with an explanation of how to use it in order to compute an answer). For the matchsticks example, the additional information given to the AR condition was "If x is the number of squares to be put in a row one can calculate the number of matchsticks, y, with the formula $y=3x+1$. (For example, if four squares are to be put in a row $y=3x+1=3\times4+1=13$ matchsticks are needed.)" Last was the question, for example "How many matchsticks do you need in order to put together one row of 6 squares of matchsticks?" The AR group was given five instances of each task set. Participants had 5 min at their disposal to solve each instance.

The version of the task designed to invite CMR did *not* include the additional information about the formula. The CMR condition was given two basic instances of each task set. Participants had 10 min at their disposal to solve each instance. After these instances they were explicitly asked to try to create a formula for the solution of the task. For the example with matches, it was asked "x denotes the number of squares in a row and y the number of matchsticks needed to create the squares. How can one describe y as a function of x? For example, it holds that $y=13$ if $x=4$ according to the figure. That y is a function of x means that there is a relationship, for example $y=14\bullet x$, $y=12-x$, $y=2\bullet x+1$, $y=3/x$ or some other relation." Participants were then given three formulas

Table 1
Content and target questions of the nine different task sets.

Content	*Target question*
Matches—one row of squares	*"How many matches do you need to put together one row of **x** squares?"*
Matches—two rows of squares	*"How many matches do you need to put together two rows of **x** squares?"*
Box—quadratic bottom	*"Compute the area of the box (the four sides and the bottom area) on a box where the edge of the bottom is **x** dm and the height is **y** dm."*
Box—equilateral	*"Compute the area of the box (the four sides and the bottom area) on a box where the edges are **x** dm."*
Box—rectangular	*"Compute the area of the box (the four sides and the bottom area) on a box where the edges of the bottom are **x** and **y** dm respectively, and the height is **z** dm"*
Flowers and stone squares—one row	*"How many stones do you need if you want to plant **x** flowers in a row?"*
Flowers and stone squares—three rows	*"How many stones do you need if you want to plant **x** flowertriplets?"*
Dots—circumference	*"How many dots make the circumference of a $\boldsymbol{x}\cdot\boldsymbol{y}$ rectangle?"*
Dots—between dots	*"How many crosses are there between the dots in a $\boldsymbol{x}\cdot\boldsymbol{y}$ rectangle?"*

as response alternatives and a fourth alternative "None of the above" (which was never correct). Participants marked which of the alternatives was the correct one and received feedback.

In sum, the key features that by design differed between the AR condition and the CMR condition were: (a) the AR condition was given a solution method while the CMR condition was not, (b) the AR condition was given five instances to practice using the formula while the CMR condition was given two instances to construct a solution method and one instance to formalize it into a mathematical function, and (c) in the CMR condition participants had 10 min at their disposal for each instance while participants in the AR condition had 5 min. The time limits were well above the average solutions times indicated by the pilot study [11].

The nine task sets were presented in the same order for each participant, starting off with an easy task and rounding off with a more difficult task, in order to mimic a realistic didactic learning scenario (i.e. the typical build-up of a mathematics text-book chapter). The ranked difficulty of a task was based on results from the previous study [11] as well as on the authors' didactical understanding of the tasks.

2.2.3. *Mathematical testing with fMRI*

Six days after the mathematical training, the participants were invited to the fMRI lab and were scanned while completing one test instance of each of the nine training task sets they had encountered, in randomized order. The test instances were identical for every participant. For the example with matchsticks, the test question was "How many matchsticks do you need in order to put together 50 squares of matchsticks in one row?" The given number in the test questions were chosen to be too large for counting the matchsticks in a simple imagined visual extension of the figure, and would instead target the knowledge of a more general solution method acquired during training. Participants were instructed to push a button on a response pad when they had solved the task and, after a variable delay, they were shown three numerical response alternatives and a fourth alternative "None of the above" (which was never correct). Participants marked which of the alternatives was the correct one and, after a variable delay, moved on to the next task (see Fig. 1 for an illustration of the scanning tasks).

The fMRI test was implemented in a blocked design, with a mathematical task always being followed by a cognitive-perceptual baseline task to allow for subtraction of perceptual, attention and reading processes (described below).

In a mathematical block (see Fig. 1A), participants were presented with a task probe for 2 s, the mathematical task question for a maximum of 30 s (self-paced), a fixation cross as an inter-stimulus interval (ISI) for a minimum of 2 s and a maximum of (32 s minus the response time: $M_{ISI}=5.2$ s), the alternative forced-choice task for a maximum of 6 s (self-paced), and an inter-trial interval (ITI) for 1.5–7.5 s. The task was self-paced in order to get measures of response times and in order to get an indication of what part of each block actually contained fMRI data related to solving the task.

The baseline task was by display perceptually very similar to the mathematical task display but unrelated to mathematics in semantic content. The task was instead to judge whether the text on the slide contained a spelling mistake or not. For example, for one baseline task the introductory text read (in Swedish, here translated to English): "Water and an engine is good to have when going by boat and there is no wind.", and was presented together with a picture of a motorboat. Instead of the mathematical question it read: "The first motorboat was built in the 1860's and was driven by coal gas, which is a kind of highly explisive town gas". Participants were instructed to push a button on a response pad when they knew the answer and, after a variable delay, they were shown two response alternatives ("Yes" and "No") and could mark what was correct. After a variable delay, they moved on to the next task. In the example above, the correct answer would be "Yes" as the text contained a spelling mistake: "explisive" instead of "explosive". In a baseline block (see Fig. 1B), participants were presented with a task probe for 2 s, the baseline task question for a maximum of 10 s (self-paced), a fixation cross as an inter-stimulus interval (ISI) for a minimum of 2 s and a maximum of (12 s minus the response time), the alternative forced-choice task for a maximum of 6 s (self-paced) and an inter-trial interval (ITI) for 1.5–7.5 s.

Upon completion of the fMRI protocol and structural images, participants filled out a follow-up questionnaire on a computer outside the scanner. Participants were asked, for each of the nine task sets, how difficult they thought it was, and were asked to respond on a scale from 1 (very easy) to 5 (very difficult).

2.3. *Image acquisition*

Image acquisition was made on a 3T GE Discovery MR 750 scanner (General Electrics). Functional T2*-weighted images were obtained with a single-shot gradient echo EPI sequence used for blood oxygen level dependent imaging. The following parameters

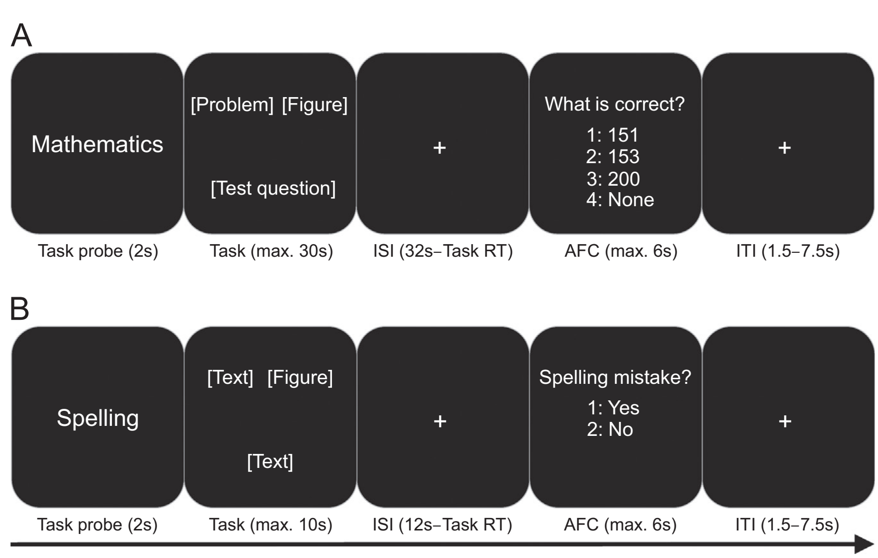

Fig. 1. Illustration of the fMRI tasks. (A) A mathematical task block. (B) A baseline task block. ISI=inter-stimulus interval; AFC=alternative forced choice; ITI=inter-trial interval.

were used for the sequence: echo time: 30 ms, repetition time: 2000 ms (37 slices acquired), flip angle: 90°, field of view: 25 × 25 cm, 96 × 96 matrix and 3.4 mm slice thickness. A 32 channel SENSE head coil was used. Signals arising from progressive saturation were eliminated through ten "dummy scans" performed prior to the image acquisition. The stimuli were presented on a computer screen that the participants viewed through a tilted mirror attached to the head coil. Presentation and reaction time data were handled by a PC running E-Prime 2.0 (Psychology Software Tools, Inc., USA) and fMRI optical response keypads (Current Designs, Inc., USA) were used to collect responses.

2.4. fMRI data analysis

The data was analyzed in SPM8 (Wellcome Department of Cognitive Neurology, UK) implemented in Matlab 7.11 (Mathworks Inc., USA). All images were corrected for slice timing, realigned to the first image volume in the series, unwarped, normalized to the standard anatomical space defined by the MNI atlas (SPM8), and smoothed using an 8.0 mm FWHM Gaussian filter kernel. Data were high-pass filtered with a cut-off of 128 s. The model consisted of two effects of interest (the mathematical task and the baseline task) and eight effects of no interest (math probe, math ISI, math alternative forced-choice, math ITI, baseline probe, baseline ISI, baseline alternative forced-choice, and baseline ITI). We chose to model the events of the two tasks as separate effects in order to incorporate potential differences during the probe, ITI, alternative forced-choice and ISI that would not be of relevance to the core question addressed. The movement parameters were included as covariates of no interest. All regressors were convolved with a hemodynamic response function. In the first level analysis, model estimations were made for each participant.

To define the regions important for the mathematical task, contrasted model estimations (Math task vs Baseline task) from each individual were taken into a second level one-sample t-test with a statistical threshold of $p < 0.05$ (FWE corrected for multiple comparisons) at the voxel level and $k > 0$ at the cluster level.

This activation map served as a mask for the between-group contrasts. Inclusive masking with SPM was used to identify whether there were condition differences in how the math-related brain regions were recruited at mathematical testing (i.e. within the defined task-related network). The condition differences were analyzed by subjecting the Math vs Baseline difference in beta values to a two-sample t-test. Because of the short duration of the intervention, we did not expect large effects of the manipulation. Hence, the statistical threshold for the masked condition contrasts was set to $p < 0.005$ (uncorrected), $k > 20$. Here it should be noted that because the search space was reduced by the masking procedure, as we sought activation differences between the two conditions in the math network, this effectively corresponded to a markedly more stringent p-threshold (see Section 3).

To investigate which regions were active during the mathematical task and played an important role for performance regardless of condition, we re-ran the second level one-sample t-test (Math vs Baseline) with performance as a covariate of interest and specifically targeted regions within the network (with the network as a mask) that covaried with performance. This analysis was also thresholded at $p < 0.005$, $k > 20$.

2.4.1. Control analyses

A number of control analyses were undertaken in relation to the condition comparisons. To investigate the possible interaction with subsample (pupils vs. students) on the condition differences, an analysis of variance was done with activity differences (Math vs. Baseline) as dependent variable and condition (AR vs. CMR) and subsample (pupils vs. students) as between-subject factors. To control for the possible effects of task difficulty on the condition differences, the procedure (the definition of brain regions important for the mathematical task to use as mask and the condition contrast within those regions) was redone using subjective difficulty ratings, averaged over the nine test tasks, and performance at test as covariates of no interest when comparing the two conditions.

3. Results

We first report behavioral results from the training and test sessions, then relate the individual difference measures to test performance, and finally report the imaging results.

3.1. Behavioral results

3.1.1. Training session

Participants spent on average 23 min (SD=6 min) solving the nine task set instances during training (excluding time for instructions and completion of follow-up questions). There was no difference in solution time between the two conditions (M_{AR}=23 min, SD=6 min vs. M_{CMR}=23 min, SD=6 min; t (1, 71)= −0.4; p=0.68) nor between the two subsamples (M_{pupils}=22 min, SD=5 min vs. $M_{students}$=23 min, SD=7 min; t (1, 71)= −1.0; p=0.32).

3.1.2. Test session

To investigate whether the CMR condition performed at a higher level than the AR condition at test, as hypothesized, while considering the possible differences between the two subsamples, we performed an analysis of variance with number of correct answers at test as the dependent variable and condition (AR vs. CMR) and subsample (pupils vs. students) as between subject factors. The results revealed a main effect of condition: the CMR condition performed better than the AR condition (F(1,69)=5.1; MSE=15.6; p=0.03: Fig. 2). There was no main effect of subsample (F(1,69)=0.07; MSE=0.23; p=0.79) nor an interaction between subsample and condition (F(1,69)=0.44; MSE=1.34; p=0.51).

Most participants needed almost all of the available solution time during scanning (30 s). Subjecting solution times (in seconds) during the test session (i.e. when the participants indicated they

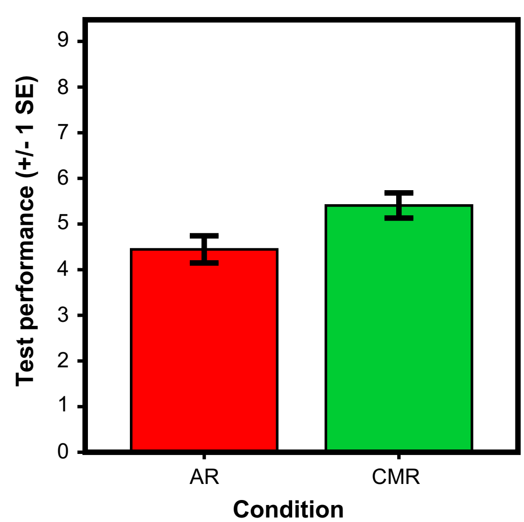

Fig. 2. Performance during the mathematical test, performed during fMRI scanning. (AR=Algorithmic reasoning group, CMR=Creative Mathematical Reasoning group). Error bars: ± 1 SE.

had solved the task) to the same type of analysis as for performance with condition and subsample as between subject factors revealed no difference between the conditions (M_{AR}=27.1 s, SD=3.5 s vs. M_{CMR}=26.5 s, SD=3.3 s), no interaction between condition and subsample but the university sample responded somewhat slower than the pupils (M_{pupils}=26.0 s, SD=3.3 s vs. $M_{students}$=27.7 s, SD=3.3 s; $F(1,69)=4.3$; $p=0.04$).

Finally, we investigated whether participants in the two conditions judged the test tasks to be equally difficult by comparing participants difficulty ratings averaged over the nine test tasks. In the AR condition the tasks were judged to be more difficult than in the CMR condition, although the difference were just bordering significant (M_{AR}=2.7, SD=0.67 vs. M_{CMR}=2.4, SD=0.82; $F(1,69)=3.8$; $p=0.06$). There was no main effect of subsample or an interaction between subsample and condition.

3.1.3. Cognitive abilities and test performance

The participant characteristics are shown in Table 2. First, as shown, the two conditions were well-matched with respect to Ravens score, mathematics grades, and gender, in both subsamples. In order to investigate which other variables except training environment (AR or CMR) had an effect on test performance, we conducted linear hierarchical regression analyses. In a first step, we included condition (i.e. training environment: AR or CMR) as independent variable and test performance as dependent variable. Condition significantly predicted performance (standardized $\beta=-0.27$, $t(71)=-2.4$, $p=0.02$). In a second step, we also included subsample (i.e., pupils or students), mathematics grades, gender, Ravens score, working memory score, vocabulary score, and ANS acuity as independent variables. This slightly modified the effect of condition on performance (standardized $\beta=-0.26$, $t(71)=-2.2$, $p=0.03$). Additionally, working memory (but not ANS acuity ($p=0.68$)) emerged as an additional significant predictor (standardized $\beta=0.30$, $t(71)=2.5$, $p=0.01$).

3.2. Imaging results

3.2.1. Definition of brain regions important for the mathematical task

First, we defined the clusters that were more involved in solving the mathematical test tasks than the baseline tasks. The contrast revealed an extensive difference in fronto-parietal brain regions, including left and right middle and superior prefrontal cortex, inferior, superior and medial parietal cortex, middle occipital cortex as well as temporal, cerebellar and subcortical regions (Fig. 3A).

Table 2
Mean (SD) participant characteristics.

Variable	AR-group	CMR-group	t-Value	p
High-school pupils				
N	19	21	–	–
Age	18.4 *(0.61)*	18.5 *(0.61)*	–	–
Math grades	16.3 *(5.49)*	15.5 *(5.22)*	–	–
Gender	7 female	9 female	–	–
Raven's	10.9 *(3.45)*	10.9 *(3.73)*	−0.009	0.99
Working memory	60.3 *(8.47)*	57.8 *(10.4)*	0.83	0.41
Vocabulary	20.0 *(4.48)*	20.0 *(3.32)*	−0.4	0.97
ANS acuity	0.21 *(0.07)*	0.23 *(0.07)*	−1.18	0.25
University students				
N	17	16	—	—
Age	20.1 *(1.35)*	19.8 *(1.11)*	—	—
Math grades	12.6 *(3.12)*	12.5 *(3.65)*	—	—
Gender	5 female	4 female	—	—
Raven's	12.1 *(4.59)*	10.7 *(2.57)*	1.05	0.30
Working memory	58.6 *(10.3)*	60.1 *(8.94)*	−0.46	0.65
Vocabulary	17.6 *(3.61)*	19.3 *(3.11)*	−1.41	0.17
ANS acuity	0.21 (0.05)	0.28 (0.19)	−1.50	0.15

Note: Math grades=range 0–20; Raven's=Raven's advanced matrices, short version; working memory=operation span; vocabulary=SRB1; ANS acuity=Weber fraction.

3.2.2. Group differences within the mathematics network

The AR > CMR contrast revealed a significant difference in left precentral cortex (*xyz:* −40 2 62, $Z=3.25$, $k=35$) and the left angular gyrus (*xyz:* −40 −66 46, $Z=2.95$, $k=52$, see Fig. 3B). The reversed contrast (CMR > AR) yielded no clusters above the statistical threshold. To further investigate the differences in angular gyrus activity, in relation to subsample, we performed an ANOVA with activity differences in angular gyrus (Math vs. Baseline) within a sphere with a radius of 5 mm around the peak voxel as dependent variable and condition (AR vs. CMR) and subsample (pupils vs. students) as between-subject factors. There was neither a main effect of subsample ($F(1,69)=2.41$; $MSE=0.40$; $p=0.13$) nor an interaction between subsample and condition ($F(1,69)=0.9$; $MSE=0.15$; $p=0.35$). Similar results were observed for left precentral cortex, with no main effect of subsample ($F(1,69)=2.34$; $MSE=0.27$; $p=0.13$) nor an interaction ($F(1,69)=1.86$; $MSE=0.22$; $p=0.18$).

As the two conditions had been found to differ in terms of test performance and perceived difficulty of the test tasks, as a control analysis we investigated whether the observed differences in brain activity between the two groups were influenced by general task difficulty (i.e. instead of being associated with task-specific processes). In so doing, we re-ran the two-sample t-test comparing AR with CMR within the brain regions defined important for mathematics, now including performance at test and perceived task difficulty (averaged over the nine test tasks) as covariates of no interest. This analysis showed that the outcome was not confounded by difficulty/performance differences, with slightly improved statistics for both angular gyrus ($\Delta z=0.71$, $k=289$) and precentral cortex ($\Delta z=0.56$, $k=91$).

3.2.3. Individual differences and mathematics

We finally examined if any of the brain regions important for mathematics played a role for performance regardless of condition. Interestingly, a cluster within the right superior parietal cortex (xyz=18 −76 58; $z=4.1$; $p<0.0001$; $k=43$) was most strongly related to performance, in that the better performance during test, the larger was the difference score in this region ($r=0.46$ between peak activation and performance: Fig. 4; $r=0.42$ after removal of a potential outlier). Activity in the right superior parietal cortex also correlated with the working memory scores ($r=0.33$; $p=0.004$). Thus, there was a correlative triad among mathematical performance, working memory capacity, and right superior parietal brain activity.

As expected, because using performance as a covariate of no interest in the group difference analysis actually improved the statistics, posthoc correlation analyses between mathematical performance and activity differences in angular gyrus and precentral cortex separately for the two groups rendered no correlations above the statistical threshold (all p's >0.005).

4. Discussion

To foster mathematical understanding is one of the most important challenges of any educational curriculum of today. To this end, teachers want to use the best teaching methods available. A dominant mathematical teaching method has been to present a solution method (e.g. a formula) and let the pupils repeatedly practice it (*algorithmic reasoning, AR*). However, the dominance of this method has been questioned (see e.g., [7,9,10]). An alternative

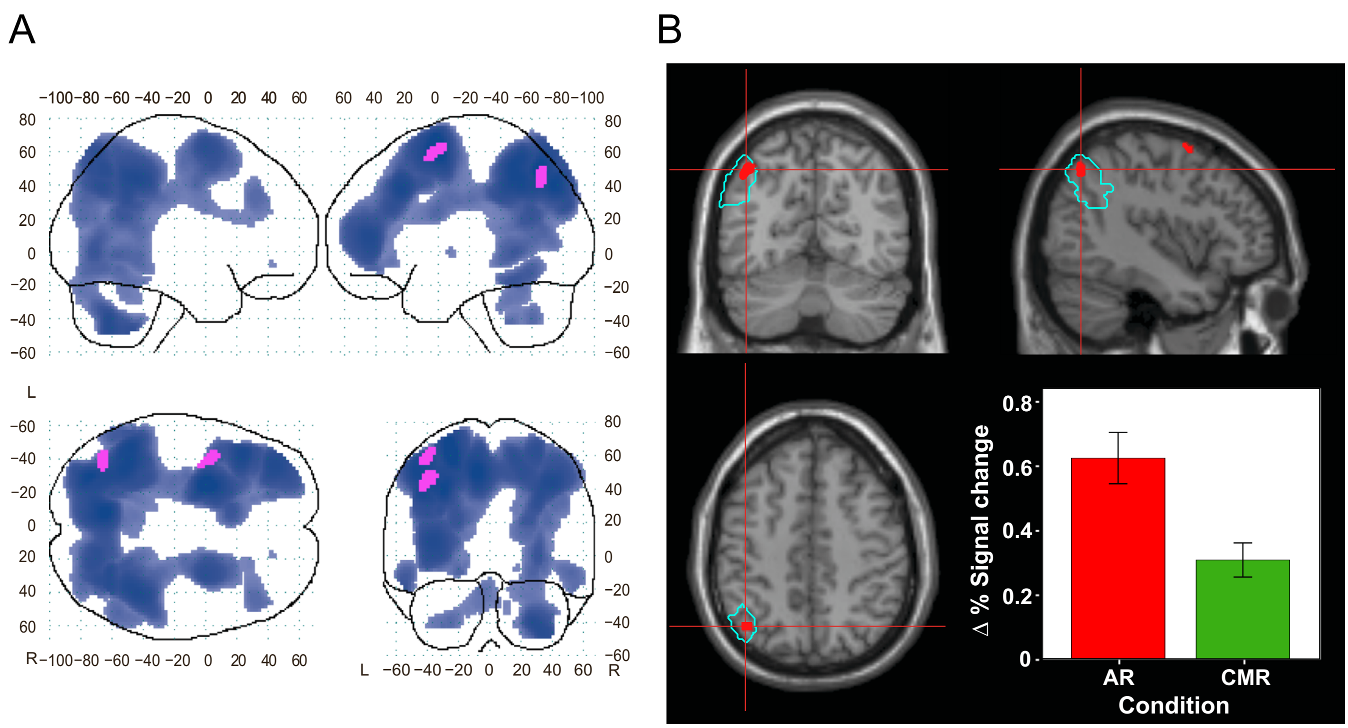

Fig. 3. (A) Blue: regions defined by the contrast Math task > baseline task (thresholded at $p < 0.05$ FWE-corrected at voxel level and $k > 0$ at cluster level). The math network included regions in fronto-parietal cortex ($xyz = -24\ 2\ 62$, $Z > 8.13$; $xyz = 30\ 0\ 62$, $Z > 8.13$; $xyz = -40\ -48\ 46$, $Z > 8.13$; $xyz = 40\ -44\ 44$, $Z > 8.13$; $k = 32{,}000$), bilateral cerebellum ($xyz = -26\ -60\ -40$, $Z = 6.31$; $xyz = 10\ -80\ -30$, $Z = 4.70$; $k = 909$), and thalamus ($xyz = -16\ -8\ 0$, $Z = 4.90$; $k = 8$). Cyan: Regions with a larger difference between AR vs. Baseline compared to CMR vs. Baseline (within mask, thresholded at $p < 0.005$ uncorrected, $k > 20$). (B) Left angular gyrus showed relatively higher activity for AR than for CMR. Bars denote difference in percent signal change (Error bars: ± 1 SE). Blue contour denotes left angular gyrus anatomically defined by the *automated anatomical labeling map* (aal: [52]) (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.).

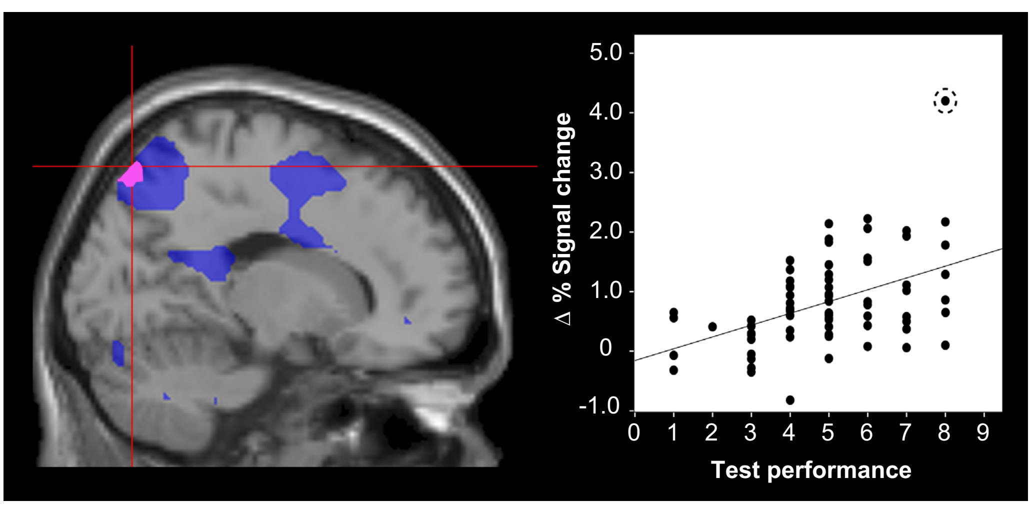

Fig. 4. Cyan: Region within superior parietal cortex ($xyz = 18\ -76\ 58$; $z = 4.1$; $p < 0.0001$; $k = 43$) demonstrating the strongest positive correlation with performance during the mathematical test (right-hand side of the figure: $r = 0.46$ between peak activation and performance). Dashed circle denote a possible outlier ($r = 0.42$ if removed). Blue: regions defined by the contrast Math task > baseline task (thresholded at $p < 0.05$ FWE-corrected at voxel level and $k > 0$ at cluster level) (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.).

method is to let the pupils create the solution method themselves (*creative mathematically founded reasoning, CMR*). In this study we compared two kinds of practice tasks – designed to promote AR and CMR, respectively – in terms of their possible different effects on mathematical performance and brain activity. We were able to demonstrate that practice tasks promoting CMR not only lead to better performance one week after mathematical training compared to practice tasks promoting AR (in line with our previous behavioral study: [11]), but also taxed brain regions important for certain mathematical component processes to a lesser extent.

4.1. Creative mathematically founded reasoning affects performance in the long-term

Interestingly, nothing in the instructions or design hindered the participants in the AR condition to reason carefully during each practice task or to try to understand whether the given solution method made sense. The AR condition was even given strictly more information in each practice instance than the CMR condition. Further, the performance advantage was observed despite the fact that the two conditions spent equal time on the training, that the

AR condition were given more than twice as many practice instances with each task set and that the conditions were matched with respect to grades and Ravens scores. This tentatively provides evidence that due to the fact that the solution method had to be self-generated with CMR, participants had an easier time accessing their knowledge of a solution method at the test one week later enabling better mathematical performance [15,16]. Further research should be devoted to clarify whether memory for the solution method indeed is superior in the CMR condition compared to the AR condition (see [11] for evidence pointing in this direction).

Where exactly lies the benefit in self-generating a solution? One general possibility is that these effects occur because the processes during self-generation overlaps those used during the subsequent testing, related to the formulation in the *transfer-appropriate processing* framework (cf., [33]; [17,18]). Another possibility is that it is cognitively more effortful to generate a solution by yourself, which in turn might strengthen the knowledge in memory and/or induce more active processing (e.g. [13,19]). In educational science it is in fact argued that "struggle" with important mathematical concepts is an important key for enhancing student's conceptual understanding [34]. The student needs to be engaged in different processes, such as generating, formulating and validating, and hence accepting the mathematical task as her own problem to solve [5]. It is thus an intriguing possibility for future research to investigate whether the CMR condition developed enhanced conceptual understanding compared to the AR condition [10].

4.2. Creative mathematically founded reasoning affects brain activity in the long-term

In line with our prediction we found that participants who had been trained in the CMR environment activated the left angular gyrus at the fMRI test session to a lower degree than did the AR condition. The relatively lower activation following CMR remained after controlling for performance differences and perceived task difficulty.

The angular gyrus is a region often implicated in imaging studies on mental arithmetic (for a review see e.g., [4]). Angular gyrus is also important for other tasks. For example, angular gyrus is implicated in models of memory-related internally directed attention [35,36] and in relation to semantic information processing more broadly, as part of a semantic control network [37,38].

In our view, this result implies that the two kinds of practice tasks (promoting AR or CMR) indeed had different effects on the fluency of memory-directed verbal processing at the one-week follow-up test [26]. Tentatively, the CMR condition had an easier time accessing their memory of a solution method, reflected in the relatively lower brain activity in left angular gyrus compared to AR.

The second region with relatively lower activity for CMR compared to AR was left precentral cortex/Brodmann area 6. This area is a region that has often been implied in neuroimaging of cognitive functions as especially related to working memory (for a review see [39]). For example, Tanaka and colleagues provided converging evidence from fMRI and TMS that activity in the lateral part of BA6 – not distant from the region reported here – is related to the updating of spatial information in working memory [40]. Precentral cortex is also commonly implicated in imaging studies of mental arithmetic in relation to untrained or more complex tasks compared to trained or less complex tasks [4]. Tentatively, this implies that participants in the CMR condition needed to engage working memory processes at test to a relatively lower degree than the AR condition (see also [11]).

4.3. Individual differences in mathematical performance

We were able to identify a large set of regions related to mathematics, regardless of condition. These regions are commonly implicated in more constrained mental arithmetic [4] as well as in problem solving more generally (see e.g., [41–43]). Within this set of regions, we observed that the right posterior superior parietal cortex played an important role for performance. Interestingly, this part of the parietal cortex has been shown to be important for tasks that require number manipulations [26]. One interpretation of the role of this region in mental arithmetic is in terms of an attentional hub, perhaps enabling internal information integration that is necessary for number manipulations [26]. Our test tasks required the manipulation of numbers in order to be solved correctly and not surprisingly, participants who activated this region more compared to the baseline task were also those that performed better.

Previous research has suggested that there might be a link between individual differences in mathematical achievement and ANS acuity [28,29], as well as working memory capacity (e.g. [27]). Here, with a regression analysis we found no support for a role of ANS acuity for mathematical performance. Interestingly, this is line with several recent studies which have also failed to establish such a link (see e.g., [44–47]. It has been suggested that, in part, this might be explained by methodological issues of the tasks used to measure ANS acuity (see e.g., [45,48–51]).

We did, however, observe a significant association between mathematical performance and working memory capacity (as measured with the Operation Span task, [31]). Performance during less constrained mathematical reasoning is thus explainable by individual variation in working memory capacity (e.g. [27]). Even though this result was not surprising given that participants were not aided by paper and pencil, or calculators, we consider this result as a reminder that each pupil's individual needs are important to consider within the realms of a teaching situation.

Finally, working memory capacity was found to correlate with activity in the right superior parietal region that was identified in relation to mathematical performance. The interrelations among these three variables (i.e. mathematical performance, working memory capacity and activity differences in right superior parietal cortex) might tentatively be seen as reflecting working memory contributions to complex mathematics, as mediated by superior parietal cortex. This might overlap the attentional contribution of parietal cortex as discussed in the model by Dehaene et al. [26] or could represent an additional independent contributor to mathematical performance.

5. Conclusion

The implications of our study are two-folded. First, our study demonstrates that different mathematical practice tasks can give rise to lasting behavioral and neural differences. *Creative mathematically founded reasoning* leads to better performance and relatively lower angular gyrus brain activity in the long-term compared to *algorithmic reasoning*. Second, these results demonstrate that the right superior parietal cortex is pivotal for mathematical performance in general, possibly reflecting attentional and/or working memory contributions to complex mathematics.

Acknowledgements

This work was supported by grants to Lars Nyberg from Umeå School of Education and the Knut and Alice Wallenberg foundation (KAW2009.0070), and to Johan Lithner from Kempe foundation (SMK-2853) and Umeå School of Education.

The authors are grateful to Mikael Stiernstedt and the staff at Umeå Center for Functional Brain Imaging (UFBI) for assistance during data collection, and to Micael Anderson and Rasmus Stenmark Persson for assistance during data analysis.

Correspondence should be addressed to Linnea Karlsson, Department of Psychology, Umeå University, 901 87 Umeå, Sweden. E-mail: linnea.karlsson@umu.se.

References

[1] Ritchie SJ, Bates TC. Enduring links from childhood mathematics and reading achievement to adult socioeconomic status. Psychol Sci 2013;24(7):1301–8.
[2] Parsons S, Bynner J. Numeracy and employment. Educ Train 1997;39(2):43–51.
[3] Rivera-Batiz FL. Quantitative literacy and the likelihood of employment among young adults in the United States. J Hum Resour 1992;27(2):313–28.
[4] Zamarian L, Ischebeck A, Delazer M. Neuroscience of learning arithmetic—evidence from brain imaging studies. Neurosci Biobehav Rev 2009;33 (6):909–25.
[5] Brousseau G. Theory of didactical situations in mathematics. Dordrecht: Kluwer Academic Publishers; 1997.
[6] Niss M. Reflections on the state of and trends in research on mathematics teaching and learning: from here to utopia. In: Lester F, editor. Second handbook of research on mathematics teaching and learning. Charlotte, NC: Information Age Publishing; 2007. p. 1293–312.
[7] Hiebert J. What research says about the NCTM Standards. In: Kilpatrick J, Martin G, Schifter D, editors. A research companion to principles and standards for school mathematics. Reston, VA: National Council of TEachers of Mathematics; 2003. p. 5–26.
[8] Bergqvist T, Lithner J. Mathematical reasoning in teachers' presentations. J Math Behav 2012;31(2):252–69.
[9] Boesen J, et al. Developing mathematical competence: from the intended to the enacted curriculum. J Math Behav 2014;33(0):72–87.
[10] Lithner J. A research framework for creative and imitative reasoning. Educ Stud Math 2008;67(3):255–76.
[11] Jonsson B, et al. Learning mathematics through algorithmic and creative reasoning. J Math Behav 2014:20–3236 2014:20–32.
[12] Slamecka N, Graf P. The generation effect: delineation of a phenomenon. J Exp Psychol Hum Learn Mem 1978;4(6):592–604.
[13] Bertsch S, et al. The generation effect: a meta-analytic review. Mem Cogn 2007;35(2):201–10.
[14] Karpicke JD, Roediger 3rd HL. The critical importance of retrieval for learning. Science 2008;319(5865):966–8.
[15] Smith R, Healy A. The time-course of the generation effect. Mem Cogn 1998;26 (1):135–42.
[16] Rittle-Johnson B, Kmicikewycz AO. When generating answers benefits arithmetic skill: the importance of prior knowledge. J Exp Child Psychol 2008;101 (1):75–81.
[17] McNamara DS, Healy AF. A procedural explanation of the generation effect for simple and difficult multiplication problems and answers. J Mem Lang 2000;43(4):652–79.
[18] Winstanley PAd. generation effects and the lack thereof: the role of transfer-appropriate processing. Memory 1996;4(1):31–48.
[19] Pyc MA, Rawson KA. Testing the retrieval effort hypothesis: does greater difficulty correctly recalling information lead to higher levels of memory? J Mem Lang 2009;60(4):437–47.
[20] Roediger HL, Karpicke JD. The power of testing memory. Perspect Psychol Sci 2006;1(3):181–210.
[21] Qin Y, et al. The change of the brain activation patterns as children learn algebra equation solving. Proc Natl Acad Sci USA 2004;101(15):5686–91.
[22] Krueger F, et al. Integral calculus problem solving: an fMRI investigation. Neuroreport 2008;19(11):1095–9.
[23] Logan GD. Toward an instance theory of automatization. Psychol Rev 1988;95 (4):492–527.
[24] Delazer M, et al. Learning complex arithmetic—an fMRI study. Cogn Brain Res 2003;18(1):76–88.
[25] Ischebeck A, et al. How specifically do we learn? Imaging the learning of multiplication and subtraction NeuroImage 2006;30(4):1365–75.
[26] Dehaene S, et al. Three parietal circuits for number processing. Cogn Neuropsychol 2003;20(3):487–506.
[27] Geary DC. Cognitive predictors of achievement growth in mathematics: a 5-year longitudinal study. Dev Psychol 2011;47(6):1539–52.
[28] Halberda J, Mazzocco MMM, Feigenson L. Individual differences in non-verbal number acuity correlate with maths achievement. Nature 2008;455 (7213):665–8.
[29] Inglis M, et al. Non-verbal number acuity correlates with symbolic mathematics achievement: but only in children. Psychon Bull Rev 2011;18 (6):1222–9.
[30] Raven J, Raven JC, Court JH. Manual for Raven's progressive matrices and vocabulary scales. Section 4, the advanced progressive matrices. Oxford: Oxford Psychologists Press; 1998.
[31] Unsworth N, et al. An automated version of the operation span task. Behav Res Methods 2005;37(3):498–505.
[32] Dureman I, Sälde H. Psykometriska och experimental-psykologiska metoder för klinisk tillämpning. Uppsala, Sverige: Almqvist & Wiksell; 1959.
[33] Morris CD, Bransford JD, Franks JJ. Levels of processing versus transfer appropriate processing. J Verb Learn Verb Behav 1977;16(5):519–33.
[34] Hiebert J, Grouws DA. The effects of classroom mathematics teaching on students' learning. In: Lester F, editor. Second handbook of research on mathematics teaching and learning. Charlotte, NC: Information Age Publishing; 2007. p. 371–404.
[35] Cabeza R, et al. The parietal cortex and episodic memory: an attentional account. Nat Rev Neurosci 2008;9(8):613–25.
[36] Ciaramelli E, Grady CL, Moscovitch M. Top-down and bottom-up attention to memory: a hypothesis (AtoM) on the role of the posterior parietal cortex in memory retrieval. Neuropsychologia 2008;46(7):1828–51.
[37] Binder JR, et al. Where is the semantic system? A critical review and meta-analysis of 120 functional neuroimaging studies Cereb Cortex 2009;19 (12):2767–96.
[38] Price CJ. The anatomy of language: a review of 100 fMRI studies published in 2009. Ann NY Acad Sci 2010;1191(1):62–88.
[39] Cabeza R, Nyberg L. Imaging cognition II: An empirical review of 275 PET and fMRI studies. J Cogn Neurosci 2000;12(1):1–47.
[40] Tanaka S, Honda M, Sadato N. Modality-specific cognitive function of medial and lateral human brodmann area 6. J Neurosci 2005;25(2):496–501.
[41] Newman SD, et al. Frontal and parietal participation in problem solving in the Tower of London: fMRI and computational modeling of planning and high-level perception. Neuropsychologia 2003;41(12):1668–82.
[42] Prabhakaran V, et al. Neural substrates of fluid reasoning: An fMRI study of neocortical activation during performance of the Raven's progressive matrices test. Cogn Psychol 1997;33(1):43–63.
[43] Reichle ED, Carpenter PA, Just MA. The neural bases of strategy and skill in sentence-picture verification. Cogn Psychol 2000;40(4):261–95.
[44] Castronovo J, Göbel SM. Impact of high mathematics education on the number sense. PLoS ONE 2012;7(4):1–16.
[45] Lindskog M, et al. Measuring acuity of the approximate number system reliably and validly: the evaluation of an adaptive test procedure. Front Psychol 2013;4:510.
[46] Szucs D, et al. Developmental dyscalculia is related to visuo-spatial memory and inhibition impairment. Cortex 2013;49(10):2674–88.
[47] Szucs D, et al. Cognitive components of a mathematical processing network in 9-year-old children. Dev Sci 2014;17(4):506–24.
[48] Gebuis T, van der Smagt False MJ. Approximations of the approximate number system? PLoS ONE 2011;6(10):9.
[49] Gilmore C, Attridge N, Inglis M. Measuring the approximate number system. Q J Exp Psychol 2011;64(11):2099–109.
[50] Price GR, et al. Nonsymbolic numerical magnitude comparison: reliability and validity of different task variants and outcome measures, and their relationship to arithmetic achievement in adults. Acta Psychol 2012;140(1):50–7.
[51] Szucs D, et al. Visual stimulus parameters seriously compromise the measurement of approximate number system acuity and comparative effects between adults and children. Front Psychol 2013:124 2013:12.
[52] Tzourio-Mazoyer N, et al. Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain. NeuroImage 2002;15(1):273–89.