Enhancing arithmetic and word problem solving skills efficiently by individualized computer-assisted practice

Wolfgang Schoppek & Maria Tulis

University of Bayreuth, Germany

wolfgang.schoppek@uni-bayreuth.de
maria.tulis@uni-bayreuth.de

Abstract

Fluency of basic arithmetical operations is a precondition for mathematical problem solving. However, training of skills plays a minor role in contemporary mathematics instruction. We propose individualization of practice as a means to improve its efficiency, so that the time spent with training of skills is minimized. As a tool to relieve teachers from the time consuming tasks of individual diagnosis, selection of problems, and immediate feedback, we have developed adaptive training software. We evaluated the application of the software in two naturalistic studies with 9 third-grade classes. Results show that even a moderate amount of individualized practice is associated with large improvements of arithmetic skills and problem solving, even after a follow-up period of three months.

2009

To appear in The Journal of Educational Research

Currently, many authors emphasize the importance of conceptual understanding for the learning of mathematics, whereas the learning of procedures is viewed as having little benefit for the development of conceptual understanding (Baroody, 2003; Fuson, Wearne, Hiebert, Murray, Human, Olivier, Carpenter, & Fennema, 1997; NCTM, 2000). However, mathematics instruction at the elementary level aims at both, conceptual understanding and computation skills. We view conceptual learning and skill development as complementary processes that stimulate each other. Conceptual understanding facilitates the development of procedures (Blöte, Klein, & Beishuizen, 2000; Carpenter, Franke, Jacobs, Fennema, & Empson, 1997; Hiebert & Wearne, 1996). On the other hand, practicing skills in order to automatize them is an important condition for reducing working memory load (Tronsky & Royer, 2002), which in turn is necessary for the construction of new conceptual knowledge (Sweller, 1988). In classes that emphasize the development of conceptual knowledge there is little time for practicing skills. Therefore, practice must be organized so as to maximize efficiency. This can be accomplished by individualizing practice sessions. The aim of this work is to investigate how much a moderate amount of individualized practice contributes to the improvement of pupils' achievements in arithmetic and mathematical problem solving. To this end, we have developed adaptive training software that supports teachers in individualizing practice.

Skill acquisition and conceptual understanding

There are numerous examples of how conceptual understanding facilitates the development of procedures (Blöte et al., 2000; Hiebert & Wearne, 1996). With a solid base of conceptual knowledge, students can invent their own procedures, resulting in a more flexible application and better transfer to novel problems (Carpenter et al., 1997). The significance of skills in the development of conceptual understanding is less obvious. In some domains such as counting and multiplication of fractions there is evidence that mastery of procedures precedes conceptual understanding (Rittle-Johnson & Siegler, 1998). But even when procedures are not necessary for the acquisition of conceptual knowledge within a domain, skills from one domain can be helpful in understanding the concepts of another domain. For example, understanding multiplication as repeated addition is cumbersome, when counting strategies for addition are still predominant (Sherin & Fuson, 2005). Additional support for this idea comes from studies showing that word problem solving performance is predicted by fluency in basic arithmetic, even after controlling for other variables such as verbal IQ and memory span (Hecht, Torgeson, Wagner, & Rashotte, 2001; Kail & Hall, 1999).

The finding that minimally guided instruction often fails to produce significant learning outcomes in lower achieving children (Kirschner, Sweller, & Clark, 2006) can also be attributed to a lack of skills. According to Kirschner et al., the activities required in instructional settings with minimal guidance make heavy demands on working memory, which impedes the acquisition of new concepts. Working memory can best be relieved by automatization of skills, which requires practice (Tronsky, 2005). On the other hand, rote learning of procedures promotes the development of buggy algorithms (Brown & VanLehn, 1981) and leads to inflexible application (Luchins, 1942; Heirdsfield & Cooper, 2002; Lovett & Anderson, 1996; Ohlsson & Rees, 1991).

How can the development of skills be supported while avoiding blind rote learning? A possible solution lies in the hierarchical structure of skills. Most complex skills are composed of simpler subskills, which can often be practiced separately. This conception, put forward by Gagne (1962), has been validated in a number of successful applications in the 1960s and 1970s (e.g. White, 1976), and has recently again become subject of debate in the "math wars" (Anderson, Reder, & Simon, 2000). Knowledge about the prerequisite relations between skills helps avoid confronting students with procedures they are not ready for. In the present work, a hypothetical hierarchy of arithmetic skills is used in the adaptive individualization mechanism of the training software (see Section 2).

Our claim for individualization is based on the theory of skill acquisition (Anderson, 1982; Anderson, Fincham, & Douglass, 1997), and on the fact that students differ in their skill development. In the first phase of skill acquisition ("cognitive phase") a declarative representation of the procedure is established and translated into behavior by slow interpretative processes. In this phase, unsupervised practice harbors the risk of students developing buggy algorithms. The next phase ("associative phase") is characterized by the proceduralization of the skill, which leads to automatization. In this phase, much practice is necessary - with no need for close supervision other than corrective feedback. Once a skill is automatic ("autonomous phase"), additional practice causes very small gains in performance. Therefore one goal of individualization is to have each student practice those skills that are in the associative phase.

To summarize, our goal is to use the hierarchical structure of arithmetic skills in order to build up complex skills gradually, guaranteeing that the learner understands each practiced subskill. This ensures that conceptual knowledge keeps pace with the development of skills, and that practicing skills means preparation for meaningful activities (Gagnon & Maccini, 2001) rather than rote drill.

Computer assisted and individualized instruction

As stated earlier, practice should be organized to maximize efficiency. Having each student in a class work on problems requiring skills that are in the associative phase of their development is efficient, because nobody is forced to practice procedures he has not yet understood nor procedures that are already automatic. Translating this into action requires diagnosing the current skill status, selecting and administering appropriate problems, and providing immediate feedback. All these are time consuming tasks, which a teacher cannot accomplish for 20+ students. Fortunately, these tasks are understood well enough to be automated in a computer program. Given the importance of computer support for individualization, the empirical research literature about the effects of computer-assisted instruction (CAI) and individualization has to be examined. Since these are independent factors, the following four combinations are possible:

a) computer-assisted, individualized instruction,

b) non computer-assisted, individualized instruction,

c) non individualized computer-assisted instruction,

d) non individualized, non computer-assisted instruction.

To our knowledge, the four conditions have never been compared in one experiment. Most studies compare interventions in accordance with Conditions a, b, or rarely c with "traditional instruction", which is often implicitly identified with Condition d. This identification is problematic, because the character of instruction in the control groups is often poorly described. Concerning Condition d, it is well documented that non individualized interventions often favor the higher achieving students (Ackerman, 1987; Helmke, 1988; Treiber, Weinert, & Groeben, 1982). On the other hand, in classes where emphasis is placed on levelling performance, higher achieving students stagnate while the gains of lower achieving students are small (Baumert, Schmitz, Roeder, & Sang 1989; Helmke, 1988; Treiber et al., 1982).

Positive effects of computer based instruction (Conditions a and c) have been found soon after computers had been available in schools (Atkinson & Fletcher, 1972; Jamison, Suppes & Wells, 1974; Mevarech & Rich, 1985). One reason for that might simply be that most students like working with computers. More importantly, compared with whole-class practice, working with computers increases the chance that every student actively solves problems, which is a contribution to valuable academic learning time (Greenwood, 1991). A recent example of individualized CAI in mathematics is the

"practical algebra tutor" (Koedinger, Anderson, Hadley, & Mark, 1997), which was designed for 9th grade, and is still very popular. It comes with a special curriculum and has proved quite successful: The effects of a one semester intervention were between d=0.3 and d=1.2 for different tests (Koedinger et al., 1997), indicating medium to large effects.

"Accelerated Math" (Renaissance Learning, Inc.) is an individualizing program that can be used with every elementary curriculum. This instructional management system helps teachers keep track of the students' progress by printing worksheets, which are filled in by the pupils, scanned, and analyzed by the computer. Studies testing implementations of "Accelerated Math" have produced mixed results: whereas Ysseldyke, Spicuzza, Kosciolek, & Boys (2003) report rather small effects of a five-months program in classes 4 and 5 between d=0.19 and d=0.40, Atkins (2005) has found detrimental effects of the use of "Accelerated math" in classes 5 through 7. Lehmann & Seeber (2005) have conducted a large scale study in 15 schools in classes 4 through 6 during a four month period. The resulting performance gains of the fourth graders were about d'=1.0 in experimental and control classes alike.

The most challenging comparison for individualized CAI is individualized instruction in small groups (Condition b). There are many examples of effective small group interventions, for example in the "cognitively guided instruction" program (Carpenter, Fennema, Franke, Levi, & Empson, 1999). In a meta-analysis about mathematics interventions for low performing children, Kroesbergen and VanLuit (2003) found that computer-assisted interventions caused smaller effects than other interventions where teachers instructed small groups[1]. We are not aware of any studies that compared computer based with non-computer based individualized practice directly. However, if both conditions produce similar effect sizes, an important argument in favor of computer-assisted practice is that personnel requirements are generally lower than for instruction in small groups.

[1] However, the control conditions in the analysed studies were quite heterogeneous: If, for example, the experimental variable consists of two variants of a CAI, smaller effects are expected than when comparing a small-group intervention with regular instruction.

To summarize, available data about the effectiveness of tools supporting individualized practice are scarce but nonetheless promising. As with PAT and "Accelerated Math", implementing the systems means a considerable intervention into the everyday routine of a school. We suspect that this is a hindrance to a wider distribution of the systems and acceptance would be greater for less invasive alternatives. With our software, we are aiming at providing such a tool that is easy to integrate within existing classroom routines. A second aspect, where we want to go beyond existing studies is to overcome the restriction to students with special needs (Kroesbergen and VanLuit, 2003). Our goal is that all students should benefit from the practice sessions.

The adaptive training software "Merlin's Math Mill"

As a tool for supporting teachers in individualizing practice of arithmetic, we developed the adaptive training software "Merlin's Math Mill" (MMM). The animated character Merlin accompanies the user through the program and provides feedback. About 4000 problems are stored in a database, together with detailed attributes. The problem selection mechanism distinguishes three basic types of problems and a number of subtypes. The basic types are "computation problems" (CP; mostly problems in the form of equations), "word problems" (WP; all types of combine, change, and compare problems, non-standard additive word problems involving ordinal numbers (Verschaffel, De Corte, & Vierstraete, 1999), multiplication and division problems, problems with more than one calculation step, "arithmetic puzzles"), and "number space problems" (NP; number line, comparison of numbers, base ten system, continuing sequences, etc.). In a first step, the algorithm determines the basic type by calculating deviations from the reference proportions of 40% CP, 40% WP, and 20% NP and selecting the type with the largest deviation. This results in a stable pattern of repeated sequences of NP-CP-WP-CP-WP.

After establishing the basic type, the subtype and the individual problems have to be determined. This step is supported by a hierarchy of problem types that was constructed on the basis of careful task analyses. Each problem type is defined by the skills that are necessary to solve it. New subskills or new combinations of subskills make up a new level of difficulty in the hierarchy. For example, compare word problems with unknown reference set are not introduced unless simpler compare problems and computation problems with the unknown at the first position have been mastered.

Skills are not confined to a certain strategy: For example, the task "crossing the tens boundary" can be performed with different strategies. To support the development of multiple strategies (Star & Rittle-Johnson, 2008), the database contains problems for each problem type that share some but not all

features. For example “crossing the tens boundary” is practiced with problems involving varying operators and placeholders. This ensures variety in the presented problems and is targeted at avoiding the development of unreflected rote strategies. Moreover, many of the problem sets contain problems of different, yet related types. Together, these measures result in the desirable shuffled format of practice with mixed problems and spaced rather than massed practice (Rohrer & Taylor, 2007).

The program creates and updates a problem type hierarchy for each user. This hierarchy is initialized based on information from the pretest. It predicts probabilities of success for each problem type using algorithms similar to those of Bayesian networks (Pearl, 2001). Adaptivity is realized on the level of sets of four to eight problems. Once selected, the problem set must be completed by the student. After completion, information about the performance is stored in the student’s individual hierarchy and can be used in the next round of selecting a problem type. Details about the hierarchy of skills, about its empirical validation, and about the selection algorithm can be found in Schoppek (2006).

Users interact with the program through a frugal interface. Figure 1 shows the screen for word problems. Several combinations of bright colors indicate the different problem types. Each problem can be tried twice. On the first incorrect answer, the character Merlin states that the solution is wrong. For word problems, Merlin also tells the users if the number or the word is wrong. On the second incorrect answer, Merlin tells the users they are wrong and shows the correct solution. In case of correct answers, Merlin responds with short statements like “correct”, “yes”, or “super”. Although the general utility of this simple type of feedback is controversial (Bangert-Drowns, Kulik, Kulik, & Morgan, 1991), we argue that it is appropriate when practicing skills that are in the associative phase of their acquisition – a condition that is provided by the problem selection algorithm of MMM.

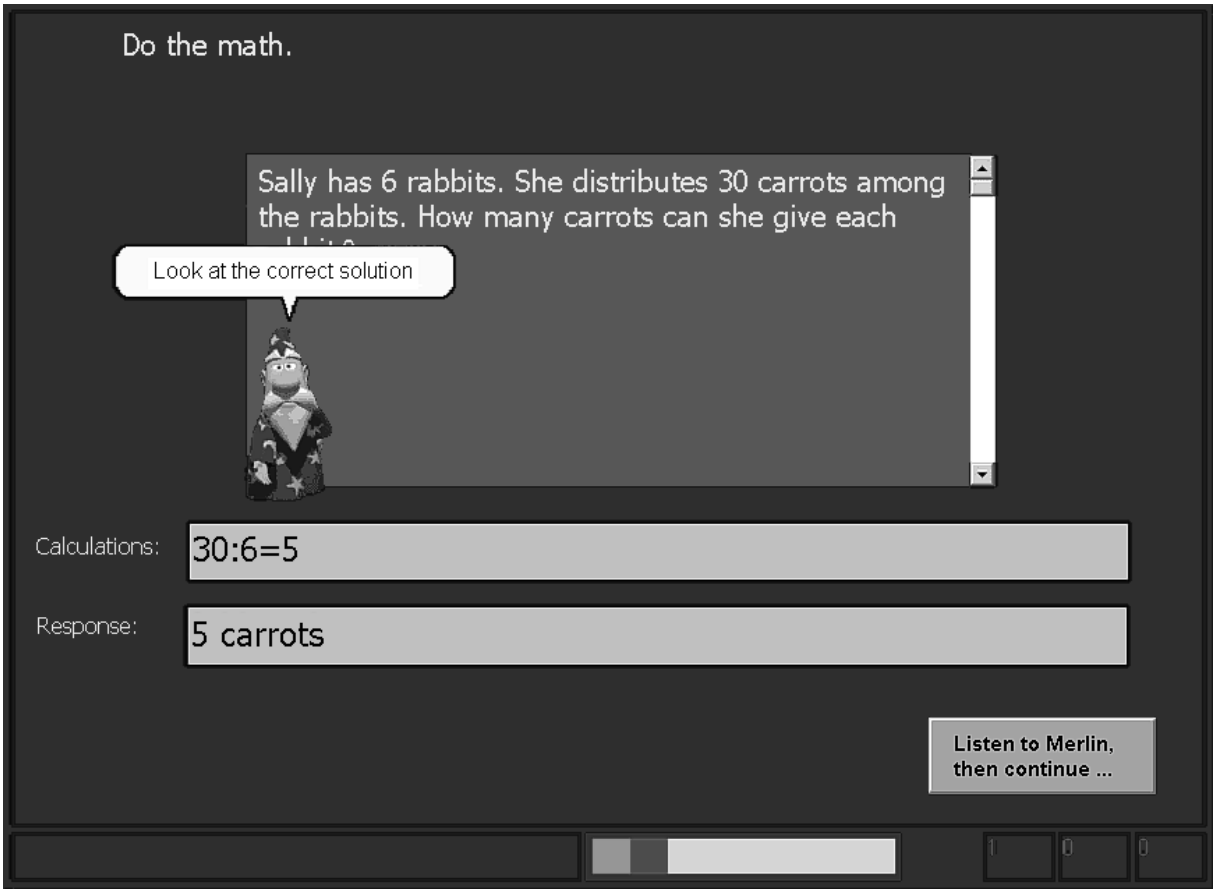

Figure 1: Screenshot of the word problem page of Merlin's Math Mill. Merlin shows the correct solution after two wrong trials. (The original software is in German.)

A bar consisting of green and red squares for correct and incorrect answers visualizes the progress within a set of problems. If the users have solved more than half of the problems of a set correctly, they can open a door in a cabinet with 40 doors. This triggers a short video clip or a joke the users can watch or read. This is meant to be a break and also an incentive to work diligently.

Pilot studies

In autumn 2004 we ran a pilot study with the first version of MMM. At that time, the program was not equipped with an automatic problem selection mechanism. The problems had to be selected manually. 20 children from two 3rd classes participated in the study on a voluntary basis. The study began in October 2004 with the pretest. In eight weeks following the pretest the 20 children had seven practice sessions of one hour each. In December the study ended with a posttest. The participants made remarkable progress from pretest (M=40.8, SD=10.8) to posttest (M=62.0, SD=11.2), which is an improvement of about two standard deviations. We believe that the automatic selection algorithm

(tested in the present studies) can hardly do better than a human expert who has access to the same information. Therefore, the gains in performance through individualized practice based on carefully hand-selected problems can serve as an upper bound estimation of what is possible by practice of that kind.

The automatic version of MMM used in the present studies was tested in another pilot study with four third graders in the summer 2005. The children were observed and interviewed about the program. This resulted in some minor modifications of the user interface and correction of bugs.

Experiment 1

According to the objective of developing a practical tool for individualizing practice, we wanted to test the effectiveness of the tool in a realistic and practicable instructional setting. That means we set up a training schedule that can be implemented without requiring above-average commitment of students or parents to learning arithmetic. We think that a weekly practice session of one hour in seven consecutive weeks meets this criterion. Our research questions were: (1) What do students gain from a small amount of additional individualized practice? (2) Do all students benefit from individualized practice with MMM in the same way? Since this was the first experiment in which the fully automatic version of MMM was employed, we wanted to test (3) how well this version worked. Concerning Question 1, we expected that trained students would improve their performance significantly more than control students, because the described individualization results in a high utilization of the limited training time. Concerning Question 2, we expected that improvements were not contingent on initial skill level. This expectation is founded on the fact that each student practices problems that match her current skills, enabling progress from any level of skills. We try to answer Question 3 by comparing effect sizes with similar studies. Specifically, we compare the effects of the automatic version with those of the pilot study from 2004, in which the practice problems were selected manually in a time consuming procedure. We consider the automation of problem selection as key to the practicability of individualized training.

Participants

IRB clearance for the study was obtained from the supervisory school authority of the city of Bayreuth, Germany. 113 children from five 3rd classes in three elementary schools in Bayreuth participated in the experiment. Parents were informed about the project with a letter distributed at school. They were

asked to indicate if their child would participate in the training sessions and return the letter with their signature. Fifty-seven children volunteered for participation, the remaining 56 served as a control group. So practical reasons prevented us from randomizing to treatment and control. Based on informal communications we found that the motivations of parents and pupils to participate or not were diverse, ranging from interest in helping low achieving children, ambitions on the side of parents, to other engagements, such as soccer training. Thus, it is not surprising that we did not find significant differences in pretest scores (t=0.90, df=108, p=.37), sex (χ^2 =2.35, df=1, p=.13), age (t=0.37, df=108, p=.71), and migration background (χ^2 =1.16, df=1, p=.28) between the groups. Although we did not randomize, we believe that the diversity of reasons to participate or not and the equivalence of the groups in important variables make it unlikely that possible training effects are mainly attributable to confounding variables.

As three pupils did not complete the posttest, the following analyses are based on N=110 participants, 57 girls and 53 boys. The mean age of the participants at pretest was 8;7 (SD=4.8). 17.5% of the participants had a migration background (i.e. the child or at least one parent has a first language other than German). Although the children with migration background scored significantly lower on the pretest than the other children (t=2.39, p<.05), migration status had no effect on the effectiveness of the training and was not involved in any significant interactions (all Fs<1.0).

Measures

We assessed arithmetic skills and mathematical problem solving with a test developed in our department. Two parts are administered with separate time constraints of 10 and 20 minutes. The first part consists of 40 problems in the form of equations. Equation problems with two operands all have a total between 21 and 100 and vary systematically in the attributes "operator" (+,-), "position of the placeholder" (first operand, second operand, total), "crossing the tens boundary" (no, yes). There are also equations with three operands with varying attributes. The second part consists of 22 word problems, including different subtypes of change and compare problems, multiplication and division problems (some of them involving remainders), and problems requiring two or more calculation steps. There are two parallel versions of the test. To avoid ceiling effects, the test contains more items than third graders can do within the given time limit. Each correctly answered equation problem scored one point. For word problems, one point was scored for the correct calculation, one point for the correct result, and one point for the correct unit stated in the answer.

The scores of the subtests "computation problems" and "word problems" are correlated with r=.63 (p<.001), justifying the calculation of a total score. The combined retest / parallel test reliability for the total score, based on the control group data is r=.82 (p<.001).

For levelling the prior skill level, we calculated residuals of the posttest total score controlled for the pretest score ("posttest residuals"). For leveling the mean skill level of the class, we pooled the variance within classes (i.e. we standardized the scores by subtracting the class mean of the pretest score and dividing the difference by the respective standard deviation). We refer to this measure as "standardized scores".

All these measures reflect the speed aspect of performance. To assess the power aspect of performance, we did separate analyses with the items that have been attempted by at least 80% of the students, applying the Rasch model. This subset of items (25 equations and 15 word problems) is consistent with a one-dimensional Rasch model and also with a two-dimensional model, assuming a second dimension for the word problems (see the results section for details). These results justify regarding the estimated person parameters for the first dimension as a measure for general arithmetic skills (referred to as "calculation") and the parameters for the word problem dimension as a specific measure for dealing with word problems (referred to as "word problem solving"). The reliabilities of the pretest, calculated with the software MULTIRA (Carstensen & Rost, 2003) using the maximum-likelihood errors of the item parameters, are .94 for "calculation" and .79 for "word problem solving".

Procedure

Participants were tested in October 2005. Students took the test during a math lesson in the classroom. In the seven weeks following the pretest, participants of the training group attended weekly practice sessions of one hour each in groups of seven to nine children. Sessions took place in a computer room of the University of Bayreuth and were supervised by one or two persons (the authors and student assistants). Supervisors helped children with handling the program and – if necessary – encouraged them to solve the problems on their own (with statements like "read the problem once more, carefully"). No specific help on the problems was provided. The posttest was scheduled in the week after the last training session. The implementation was the same as for the pretest, except that each participant got the other version of the test. After scoring the tests, pupils and teachers were debriefed about the results.

Data analysis

There are two aspects of training effects: Differences between the conditions and improvement. To test if the trained groups outperformed the control groups after the practice sessions, we conducted multivariate between-subjects analyses of covariance on the posttest scores in computation problems and word problems. Although the two groups did not differ in their pretest scores, we selected them as covariates to increase power by reducing error variance and to control for individual differences in initial arithmetic skills. We entered class as a factor to test possible class effects. Where the emphasis was on different improvements of subgroups with high vs. low initial skill level, we calculated analyses of variance with the total scores in pretest and posttest as repeated measures.

Effect sizes are being expressed as partial η^2-coefficients, which represent the proportion of variance explained by the respective factor or interaction. According to Cohen (1988), η^2–values between .05 and .12 are medium effects; η^2–values greater than .12 are large effects. With N=110 and α=.05 the analysis of variance detects medium effects with a power of 1-β = .74 and large effects with a power of 1-β = .99.

Results

We begin the presentation of the results with a more qualitative description of the training sessions, followed by quantitative analyses of the progress the different groups made from pretest to posttest.

Description of the training sessions

In the seven training sessions, the children worked through 54 to 168 computation problems (M=104), 64 to 145 word problems (M=102), and 24 to 73 number space problems (M=48). The large range of number of finished problems raises the question about individual differences in motivation. Of the five participants who solved the least number of problems, one boy missed a training session, but was inconspicuous otherwise. Two girls and a boy – all with a pretest performance above average – spent unreasonably much time with certain problem sets, particularly with word problems. One of the girls had to be admonished repeatedly not to chat with her neighbor. These children seemed to lack motivation to solve the practice problems; accordingly, the two girls gained but two points, the boy ten points from pretest to posttest. Another girl – with a pretest score below average – worked slowly, but thoroughly throughout the training period, showing no signs of lacking motivation. She gained 23 points in the posttest.

A few participants tried to skip problems they did not like by clicking away the entire set without reading the problems (which requires two clicks per problem plus some time to wait for Merlin finishing his comments). This was particularly the case with word problems, where ten children clicked away more than 10% of their word problems. However, in total only 5.5% of the word problems were clicked away. For the computation problems and the number space problems the proportions of clicked away problems were 2.3% and 3.4% respectively. The high proportions of finished problems indicate that most children were highly motivated to work with the program. This is also supported by the comments of children, many of which told us they liked the number space problems and the computation problems, and by the statements of all teachers that the children were looking forward to attending the training sessions. We observed that the children were particularly eager to watch the videos. Occasionally, some children left their seats for watching another child's video. In these cases, the children were admonished by the experimenter and returned to their seats.

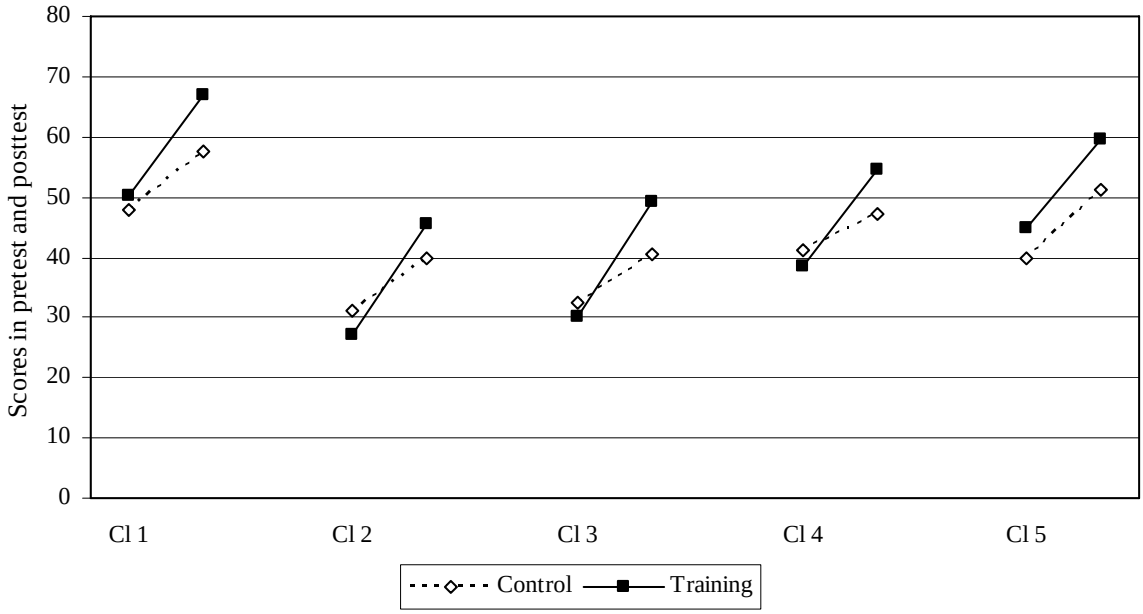

Figure 2: Improvements from pretest to posttest in five classes of Experiment 1

Differences between the groups

Figure 2 shows the means of pretest and posttest total scores in the trained and untrained groups for each of the five classes. It is obvious that although pretest levels differ between classes, the trained children show more improvement than the untrained children in all classes. To examine effects of the

training on computation problems and word problems, we calculated a multivariate analysis of covariance (MANCOVA) with the posttest scores in the two kinds of problems as dependent variables, treatment and class as between-subjects factors, and pretest scores in computation and word problems as covariates. Our hypothesis that trained groups outperform untrained groups at posttest would be supported by a significant main effect of treatment. Class effects on improvement can be judged by the interaction between treatment and class. Means, standard errors, *F* statistics, and effect sizes of this analysis are summarized in Table 1.

Table 1: Means and standard errors of total scores and subscores in Experiment 1

	Condition						
	Control (n = 56)		Training (n = 54)				
Measure	*M*	*SE*	*M*	*SE*	*F*(*df*s)	η^2	*d*
Pretest							
Total	40.1	2.3	37.4	2.3	0.12 (2, 99)	<.01	
CP	14.3	0.9	13.7	0.9	0.02 (1, 100)	<.01	
WP	25.8	1.6	23.7	1.4	0.10 (1, 100)	<.01	
Posttest							
Total	49.1	2.4	54.7	2.1			
CP	16.6	0.9	18.2	0.9			
WP	32.5	1.7	36.4	1.4			
Adjusted Posttest[a]							
Total	48.1	1.2	56.1	1.2	10.87 (2, 97)***[b]	.18	0.78
CP	16.3	0.5	18.6	0.6	7.81 (1, 98)**	.07	
WP	31.6	0.9	37.3	0.9	20.09 (1, 98)***	.17	

[a] Mean adjusted pretest scores are for Total: 38.8, CP: 14.0, WP: 24.8
[b] *: p<.05, **: p<.01, ***: p<.001

The MANCOVA revealed significant main effects of treatment (F(2, 97)=10.87, p<.001, η^2=.18) and of both covariates (CP: F(2, 97)=24.51, p<.001, η^2=.34, WP: F(2, 97)=38.82, p<.001, η^2=.45), but no main effect of class (F(8, 196)=0.63, p=.754, η^2=.03). This means that the trained groups outperformed the control groups in the posttest, and that the overall improvements were about the same in all classes. The interaction between treatment and class is not significant (F(8, 196)=0.62, p=.766, η^2=.02), indicating that differences between trained and untrained groups did not differ between classes.

Univariate analyses showed that the differences between trained and untrained children were significant for both dependent variables: computation problems (F(1, 98)=7.81, p<.01, η^2=.07), and word

problems ($F(1, 98)=20.09$, $p<.001$, $\eta^2=.17$). Overall, the training resulted in a large effect of $\eta^2=.18$, which amounts to a d of 0.78 (calculated using "posttest residuals").

Role of initial skill level

One aim of individualized practice is supporting students at all performance levels equally well. To test if practicing with MMM meets this requirement, we split each participating class at the median of the pretest total score and used the membership in the lower vs. upper half as factor "ability level" in an ANOVA with repeated measures on pretest and posttest, together with the factor treatment. Since in preliminary analyses class was not involved in any significant interactions, we didn't enter this factor in the analysis. Note that for the ANOVA with repeated measures it is not the main effect of treatment that is critical for demonstrating the training effect, but the interaction between time and treatment. Differences in training effects between the lower and upper half would result in a significant three-way interaction between time, treatment, and ability level.

As expected, this interaction was not significant ($F(1,100)=1.50$, $p=.22$, $\eta^2=.01$). So we can conclude that higher as well as lower achieving students benefit from practicing with MMM in about the same way. In the control group, both ability levels had a mean improvement of nine points. In the trained group, the lower half improved their mean test score twenty points, the upper half fifteen points.

Rasch analyses

Rasch analyses of the test data allow more detailed analyses of the training effects. As mentioned earlier, the data are consistent with a two-dimensional Rasch model, assuming a dimension "calculation" for all problems and a second dimension "word-problem-solving" for the word problems only. The estimated person parameters in the two dimensions reflect calculation and word problem solving skills more purely than the raw scores. According to Rost (2004), estimated person parameters can be used to assess trainings effects.

Model fit was tested using the bootstrap method, where 100 samples were simulated using the estimated item parameters, and the χ^2-statistics of the real sample was compared with the distribution of these statistics in the simulated samples. The empirical χ^2-statistics indicate the significance of the deviation between data and model. If these statistics lie within the left 95% of the distribution of the same statistics in the simulated sample, the model is accepted as fitting the data. For the pretest data, the empirical χ^2 has a rank of 86 in the bootstrap sample, meaning that there are still 14 % simulated cases with larger

χ^2. For the posttest data, this rank is 71. Hence the deviation of the data from the two-dimensional Rasch model can be attributed to chance.

Like for the raw scores, we calculated a MANCOVA with the factors treatment and class, and the Rasch parameters for the dimensions "calculation" and "word-problem-solving" as covariates (pretest) and dependent variables (posttest). Again, we found a significant main effect of treatment ($F(2, 97)=4.13$, $p<.05$, $\eta^2=.08$), and of the covariates (both Fs$(2, 97)>17.0$, $p<.001$). No other effects are significant. Particularly, class has virtually no effect ($F(8, 196)=0.55$, $p=.82$, $\eta^2=.02$). Thus, confirming the result of the analysis using the raw scores, we can conclude that the effect of the training does not depend on different performance levels in the classes. This indicates that the practice provided by MMM is really adaptive. However, the effect size for the Rasch parameters ($\eta^2=.08$) is lower than for the raw scores ($\eta^2=.18$). We interpret this as hint that the training affects the speed aspect of performance more than the power aspect.

Univariate tests revealed that the differences between the groups are mainly due to the "calculation" parameter ($F(1, 98)=4.65$, $p<.05$, $\eta^2=.05$). The effect for "word-problem solving" is not significant ($F(1, 98)=0.08$, $p=.778$, $\eta^2<.01$). How can we resolve this seeming discrepancy with the results for the raw score of solved word problems? Remember that trained students solved more word problems in the posttest than untrained students. One explanation could be that the improvement in solving word problems must be attributed to the gains in "calculation" and the training contributed nothing to the specific skill of constructing a mathematical model from the text of a word problem. This is disappointing considering our emphasis on word problems and the fact that pupils have spent an average of 3 h, 23 min working on word problems across the seven practice sessions (compared to 47 min working on computation problems).

Comparison with the pilot study

Since in the 2004 pilot study most conditions, like the practiced problems, the age of the subjects, the school, the training procedure, the tests, and the pretest results were very similar to those in Experiment 1 (pilot study: $M=40.8$, $SD=10.8$; Exp.1: $M=38.8$, $SD=16.1$), we can compare the improvements. The results of the standardized posttest scores for each class are shown in Figure 3. Whereas in the pilot study participants gained about two standard deviations from pretest to posttest, in Experiment 1 this was between 1.1 and 1.5 standard deviations. Subjects in the control condition consistently showed smaller improvements below 0.8 standard deviations.

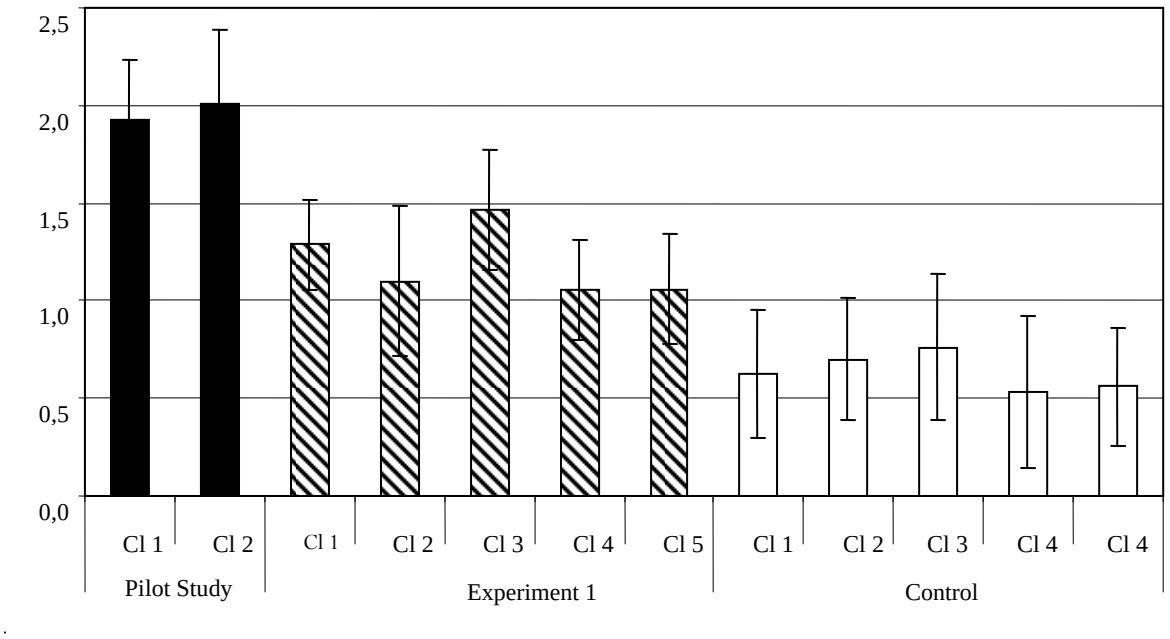

Figure 3: Standardized posttest scores in two classes of the pilot study and five classes of Experiment 1

Discussion

Since we did not randomize the subjects to the conditions in Experiment 1, the following conclusions have to be stated with the reservation that the observed effects could be attributed to some confounding variable. Although we judge it to be unlikely that volunteer status alone could explain the large training effects, the results need to be replicated in a more controlled experiment.

With these caveats in mind we could show that a small amount of individualized practice with MMM can stimulate considerable gains in arithmetic performance. This is particularly true for the condition where practice problems were selected manually. In view of the often replicated finding that algorithms are better than humans in many diagnostic tasks, especially in medical diagnosis (e.g. Bassoe, 1995; Forsstrom & Dalton, 1995), this is surprising. We attribute our result to the facts that the same experts who devised the algorithm also selected the problems and that the algorithm cannot handle exceptions as well as the experts. For example, when a child performs unusually well on multiplication problems, but still poorly on addition problems, the expert would postpone more multiplication problems and focus on addition for a while, whereas the algorithm would progress on the multiplication strand of the hierarchy. We take these findings as incentive to improve MMM's problem selection algorithm.

Low performing students gained at least as much from practice with MMM as high performing students. In the light of findings that traditional practice usually supports either group while neglecting the other (Ackerman, 1987; Helmke, 1988), this is a desirable outcome.

The Rasch analyses suggest that the observed improvements in solving word problems should be attributed to gains in computation skills. This is consistent with findings that basic arithmetic skills explain unique variance in word problem solving performance (Hecht et al., 2001; Kail & Hall, 1999). Probably, the trained children solved word problems just faster. Considering that we aim to enhance children's problem solving skills by practicing with MMM, we need to improve the way in which word problems are selected and presented.

Experiment 2

In Experiment 1 we tested the potential of individualized training administered in a feasible amount of 1 hr per week. The aim of Experiment 2 was to test the effects of individualized practice in the classroom. We scheduled one weekly lesson (45 min) of working with MMM during regular math instruction and compared the improvements with classes who received traditional instruction. This offered us a control condition that involved the same amount of training as the experimental condition. Our hypothesis was that replacing parts of the mathematics lessons with individualized practice would induce a measurable difference in improvement of pupils' arithmetic and problem solving skills. We expect this because the treatment differs from traditional instruction in two aspects. First, practicing with the computer increases students' academic engagement rate, which as part of academic learning time is an important condition for learning (Brophy, 1986). Second, the design involves splitting of the class such that the teacher spends two lessons per week with only half of the class, allowing for more specific instructional measures. As in Experiment 1, we expected the intervention to be equally effective for low and high achieving students.

Participants

Four 3rd classes from two schools in the region of Bayreuth, Germany participated in the experiment. Both schools were located in non-urban communities. IRB clearance was obtained from the supervisory school authority of the district of Oberfranken, Germany. Parents were informed beforehand about the project with a letter distributed at school. Since the tests and the training sessions were part of the math classes, no specific permission from the parents was necessary. All 94 children, 51 boys and 43 girls of

the four classes participated. The mean age of the participants at pretest was 9;1 (SD=4.7). There were no significant differences between the four classes in sex (χ^2 =1.73, df=1, p=.19), age (F(3,90)=1.09, p=.36, η^2=.04), and migration background (χ^2 =0.10, df=1, p=.53). However, one of the control classes had a significantly lower average pretest score (F(3,90)=7.89, p<.01, η^2=.21, Scheffé procedure, p<.05) than all other classes.

Measures

We used the same test as in Experiment 1 to assess arithmetic skills and mathematical problem solving. As a follow-up test, we used the DEMAT 3+ (Roick, Gölitz, & Hasselhorn, 2004). This test has three subscales, two of which (arithmetic and word problems) are expected to be sensitive to the training. Therefore, we combined the subscales arithmetic and word problems by summing up their scores. The third scale, geometry, should not differ between the treatments. For comparing all measures in one analysis, we calculated standardized residuals of the posttest and follow-up test scores controlled for the pretest score. Two children were absent at the posttest; two other children were absent at the follow-up test. We replaced the missing values with the respective class means.

Design and procedure

The experimental design involved a treatment factor (individualized practice vs. regular instruction), and a repeated measures factor with the levels pretest, posttest, and follow-up test. Classes were randomly assigned to treatment vs. control conditions. In two classes, children practiced with MMM, the other two classes served as control classes. Because of the small number of classes we calculated all analyses also with the treatment factor replaced by the factor class. As the alternative analyses yielded equivalent results, we report only the effects of the treatment factor.

All classes had their pretest in April 2006. In the nine following weeks, the subjects in the experimental condition practiced 7 times 45 min with MMM. The nine weeks included two weeks of vacation. For the practice sessions, each class was divided into two halves according to the pretest scores. While one half worked with MMM, supervised by the first author or a student assistant, the other half stayed with the teacher and vice versa. Conditions during the sessions were as described in Experiment 1. In July 2006, all classes were tested again. As in Experiment 1, each participant got the version of the test she hasn't had in the pretest. The follow-up test was administered in October 2006. At the end of the study, teachers were handed out the results of the tests and were told to debrief the pupils.

Results

Similar to Experiment 1, the children worked through 65 to 178 computation problems (M=110), 47 to 132 word problems (M=93), and 33 to 92 number space problems (M=59). The problem with clicking away word problems was more serious than in Experiment 1: On average, 20% of the word problems were clicked away! Since the problems that are clicked away are marked as not solved, the selection algorithm brings up these problems again later. Thus, many of the problems that had been clicked away at first were solved at the next opportunity. About one quarter of the children expressed their annoyance at the word problems, particularly during the last two sessions, when the weather was hot. However, the absolute number of finished problems, which was not lower than in Experiment 1, and many positive comments indicate that again most children were motivated to work with the program. The aspiration to watching the videos was the same as in Experiment 1.

As can be seen in Table 2, the control group scored significantly lower in the pretest than the trained group (which is due to only one of the control classes). Therefore, controlling for pretest scores is important in the following analyses.

Table 2: Means and standard errors of total scores and subscores in Experiment 2

	Condition						
	Control (n = 58)		Training (n = 36)				
Measure	M	SE	M	SE	$F(df\text{s})$	η^2	d
Pretest							
Total	44.4	2.5	55.4	3.0	5.30 (2, 91)** [b]	.10	
CP	14.6	0.9	19.6	1.4	10.36 (1, 92)**	.10	
WP	29.8	1.6	35.8	2.0	5.45 (1, 92)*	.06	
Posttest							
Total	49.2	2.6	65.1	3.0			
CP	17.0	1.0	21.2	1.5			
WP	32.2	1.7	43.9	1.9			
Adjusted Posttest[a]							
Total	53.1	1.1	58.8	1.4	9.37 (2, 89)***	.17	0.65
CP	18.9	0.5	18.1	0.9	0.68 (1, 90)	<.01	
WP	34.5	0.9	40.3	1.1	15.52 (1, 90) ***	.15	

[a] Mean adjusted pretest scores are for Total: 48.6, CP: 16.5, WP: 32.1
[b] *: p<.05, **: p<.01, ***: p<.001

To test the effects of individualized practice on computation problems and word problems, we calculated a MANCOVA with treatment as between-subjects factor, pretest scores in computation problems and word problems as covariates, and the respective posttest scores as dependent variables. As expected, we found a significant main effect of treatment (F(2, 89)=9.37, p<.001, η^2=.17), and of the

covariates (CP: $F(2, 89)=50.08$, $p<.001$, $\eta^2=.53$, WP: $F(2, 89)=18.04$, $p<.001$, $\eta^2=.29$). Univariate analyses showed that, deviating from the results of Experiment 1, only word problems contributed significantly to the treatment effect ($F(1, 90)=15.52$, $p<.001$, $\eta^2=.15$), whereas calculation problems did not ($F(1, 90)=0.68$, $p=.412$, $\eta^2<.01$). As in the previous experiment, the training resulted in a large effect of $\eta^2=.17$; however, $d = 0.65$ (calculated using "posttest residuals") indicates a medium effect.

To test if students at all levels of achievement profited from the training, we did the same analysis as in Experiment 1: We split each class at the median of the pretest total score and used the membership in the upper vs. lower half as an independent variable in an ANOVA, together with the factors time (pretest vs. posttest) and treatment. The fact that we found no three-way interaction ($F(1,92)=0.62$, $p=.44$, $\eta^2=.01$) indicates that the gains of lower achieving students and higher achieving students are similar. In the control groups, both ability levels had a mean improvement of five points. In the trained groups, the lower half increased their mean test score about nine points, the upper half gained eight points.

Standardized scores, using class specific means and standard deviations of the pretest total score as transformation parameters, are presented in Table 3 (means and standard deviations of the pretest values are all 0.0 and 1.0 due to the standardization procedure). At the class level, although Class 1 and Class 4 have their means very close, the ranking of the four means is as expected, with the experimental classes showing larger gains than the control classes. Interestingly, Class 4 is the one with significantly lower pretest scores than all other classes.

Table 3: Means and standard deviations of standardized total scores in posttest of Experiment 2

	Class 1 Treatment	Class 2 Treatment	Class 3 Control	Class 4 Control
M	0.48	0.70	0.26	0.43
SD	1.03	0.94	1.09	1.12

Follow-up test

The means in Table 3 represent average gains in units of the pretest standard deviation. Compared with the results of Experiment 1 (see Figure 3), these values are rather low. To a certain extent, this is

consistent with earlier findings of our group where gains were lower in the second half of the school year than in the first half (Laue & Putz-Osterloh, 2002). An unusual hot summer and the FIFA world cup that took place in Germany during the last sessions and the posttest may also have contributed to the low gains. Therefore, the follow-up test, administered in October, should corroborate the differences between the treatments.

To test these differences, we ran a MANCOVA with the factor treatment, pretest scores in computation problems and word problems as covariates, and the three subscales of the DEMAT 3+ as dependent variables. The three subscales are "computation problems", "word problems", and "geometry". Since geometry was not subject of the training, the geometry scale can be used to test the specificity of the training: No differences between the trained classes and the control classes are expected in that scale.

The only significant effect is the main effect of treatment ($F(3, 88)=3.39$, $p<.05$, $\eta^2=.10$). The means indicate that the difference between experimental and control groups has not vanished, even three months later: The size of the effect, calculated using residuals of the follow-up scores in computation problems and word problems is still $d=0.65$. Univariate analyses confirm the expectation that there are no differences in the geometry scale ($F(1, 90)=0.68$, $p=.413$, $\eta^2<.01$), but only in the computation problems ($F(1, 90)=8.47$, $p<.01$, $\eta^2=.09$), and in the word problems ($F(1, 90)=4.73$, $p<.05$, $\eta^2=.05$).

Rasch Analyses

To tease apart the potential influence of the treatment on doing calculation vs. solving word problems, we did the same Rasch analyses as in Experiment 1. Model fit was calculated with MULTIRA, using the bootstrap method indicating only non-significant deviations of pre- and posttest data from a two-dimensional model with two dimensions: "Calculation" for all problems and "word problem solving" for the word problems only.

A MANCOVA with the between-subjects factor treatment and the posttest Rasch parameters "calculation" and "word-problem-solving" as dependent variables, controlling for the same parameters at pretest, resulted in a significant main effect of treatment ($F(2, 89)=5.45$, $p<.01$, $\eta^2=.11$) and of the covariates ("calculation": $F(2, 89)=126.5$, $p<.001$, $\eta^2=.74$; "word problem solving": $F(2, 89)=11.07$, $p<.001$, $\eta^2=.20$). Univariate tests revealed no significant effect of treatment on the parameter "calculation" ($F(1, 90)=2.01$, $p=.160$, $\eta^2=.02$). For "word problem solving", however, the effect of

treatment is significant ($F(1, 90)=10.61$, $p<.01$, $\eta^2=.11$). This means that in Experiment 2, the training had a specific effect on solving word problems, whereas in Experiment 1 the training exerted its effects mostly on the "calculation" dimension.

Discussion

Experiment 2 shows that individualized practice with MMM supports the development of arithmetic skills even when employed during regular instruction (i.e. the sessions were not additional practice opportunities, but replaced traditional math lessons). This effect lasted at least three months as demonstrated by the follow-up test, which yielded the remarkable effect size of $d=0.65$. The results indicate that traditional instruction does not offer enough opportunities for efficient practice.

Different from Experiment 1, pupils in Experiment 2 have made progress specifically in solving word problems. We had not expected this effect. An explanation a posteriori could be that teachers had used the lessons they spent with only half of the pupils for providing deepened and differentiated instruction about solving word problems. As we have not collected data about the teachers' instruction during the MMM sessions, we must postpone testing this hypothesis in a future study.

General discussion

Two studies have been conducted to investigate what students gain from a small amount of individualized practice using the adaptive practice software "Merlin's Math Mill". As both studies have their methodological limitations (which are quite common in field research), we cannot draw definite conclusions. There are several arguments in favor of the interpretation that the observed effects are caused by the interventions and not by confounding factors. First, both studies yielded the predicted results under different circumstances. It is very unlikely that in Experiment 1 the effects were caused by the volunteer status *and* in Experiment 2, where volunteer status was held constant the effects were caused by unspecific properties of the classes. (Recall that in Experiment 1 class was not involved in any significant interactions). Second, in both Experiments the trained group and the control group were equivalent in important variables. Third, in Experiment 2, we found that performance at follow-up differed only in the trained areas between the conditions, whereas performance was equal in the untrained area of geometry. Unspecific factors would likely have affected all these areas equally. Nevertheless, the present findings should be replicated under controlled experimental conditions.

The reported studies have demonstrated that a moderate amount of computer-assisted individualized practice, based on our hypothetical hierarchy of skills, can improve students' mathematics performance considerably. This indicates indirectly that the children don't have sufficient opportunities for practicing relevant skills. It is important to note that the results were found with a small intervention of seven weekly hours of practice. This means that practicing with a tool like MMM is indeed efficient. It also means that such a tool can be easily integrated within existing classroom routines, and it can be combined with other approaches to teaching mathematics as for example the conceptual or the problem solving approach (Baroody, 2003). We expressly do not claim that elementary mathematics instruction should be entirely computerized. Accordingly, MMM was not designed as a tool for learning new matter.

Both experiments have shown that low achieving and high achieving students gained equally from practicing with MMM. This important finding suggests that the tool could be useful for supporting children with learning disabilities. In our samples, many children with very low pretest scores (9-12 points) made progress in the range of 20 points. This supports the view that the hierarchy of skills, which forms the basis for the problem selection mechanism, is a suitable model for skill development on all levels of expertise.

Compared to studies that employed similar programs, practicing with MMM has yielded considerable effect sizes: In Experiment 1, the difference in total posttest scores was d=0.78, in Experiment 2, this measure was d=0.65. This is well above the average effect size of d=0.35 that Kulik (1994) calculated in his meta-analysis about CAI and more than found with applications of "Accelerated Math" (Ysseldyke et al, 2003; Lehmann & Seeber, 2005). Uniform practice (where all students work on the same problems) on computers may even have no effect at all. In a study by Desoete, Roeyers, and De Clercq (2003), such a condition was used in one of four comparison groups for a metacognitive training. It produced no larger gains than the control condition.

We attribute the success of our intervention mainly to the individualization. Each student was given the opportunity to practice those skills that were in the associative phase of their development, resulting in large speed gains together with a low risk of developing buggy algorithms (Brown & VanLehn, 1981). The reason for this is that skills in the associative phase have been installed in principle, but not automatized. As a consequence, gains in speed and fluency can be achieved without elaborated feedback or additional instructional support. This view is supported by the finding that effect sizes for speed measures were larger than those for power measures. Although speeding up skills is not a central goal of

mathematics instruction, one should keep in mind that it is associated with progressive automatization, which in turn sets working memory capacity free for more difficult problems (Tronsky & Royer, 2002). Another consequence of individualization is that each student works on problems of moderate difficulty, which is the condition for maintaining motivation because the problems are neither boring because of being too simple nor frustratingly difficult. Informal feedback collected from the pupils confirmed that most of them enjoyed working with the software, with one exception: Some did not like the word problems and tried to avoid them.

The impact of our interventions on word problem solving was – although significant – smaller than expected, confirming findings that small interventions have no or even adverse effects (Elia, Gagatsis, & Demetriou, 2007; Van Essen & Hamaker, 1990), and large effects can only be obtained through large instructional interventions (e.g. successful applications of schema based instruction: Fuchs, Fuchs, Prentice, Hamlett, Finelli, & Courey, 2004; Jitendra, Griffin, Haria, Leh, Adams, & Kaduvettoor, 2007). Compared with these studies, our training software provides a greater variety of problem types, including multiplication, division, and multistep problems. If this is a strength or a weakness is to be determined in future studies. In any case, the rationale of selecting and administering word problems in MMM should be reconsidered: We expect to make word problems more attractive by reducing the proportion of these problems, so that the times spent with word problems and with computation problems are more balanced, and by personalizing the texts (Ku, Harter, Liu, Thompson, & Cheng, 2007).

While both interventions – additional practice and practice during class – generated considerable effects, we prefer the latter because of its greater potential for individualization: In this arrangement, the teacher spends two lessons per week with only half of the class. This time can be used for recapitulation, for introducing advanced matter, or for specific compensatory measures.

Finally, we want to discuss potential issues of our concept of individualized practice. There is an inherent danger in practicing procedures that students apply them rigidly (e.g. in situations where other solutions would be more appropriate) or blindly (without understanding). We tried to avoid this by varying systematically the contexts in which each subskill appears (Stark, Mandl, Gruber, & Renkl, 1999), for example in computation problems, in several types of word problems, and in arithmetic puzzles. About 80% of the sets of word problems contained at least two subtypes, such as combine and change problems. However, one-step addition or subtraction problems were not mixed with one-step multiplication or division problems within one set. To foster problem solving flexibility, we plan to

introduce more mixed problem sets in future versions of the software. Rote learning of procedures is associated with poor performance in transfer problems (Hilgard, Irvine, & Whipple, 1953; Paas & Merriënboer, 1994). Due to the fact that each of our participants has practiced a unique sequence of problems, the distance of transfer between the practiced problems and the posttest problems is different for each of them. Therefore we cannot test directly how well the participants solved transfer problems.

Another possible issue is the fact that students cannot select the problems themselves, which contradicts the pedagogic principle of supporting self-determination. However, most third graders have not developed enough metacognitive skills to estimate the difficulties of problems (Desoete & Roeyers, 2006), and tend to select easy problems. In a current study, we are investigating the effects of different selection modes. As we believe that guided practice is just one element of good mathematics instruction, we think that self-determination should better be supported in other activities, such as exploration.

Acknowledgments

We thank Claudia Schenk and Lisa Priester for their help in conducting the studies, David Peebles and Richard Young for their help with language related issues, and Wiebke Putz-Osterloh and three anonymous reviewers for their valuable comments.

References

Ackerman, P. L. (1987). Individual differences in skill learning: An integration of psychometric and information processing perspectives. *Psychological Bulletin*, *102*, 3-27.

Anderson, J. R. (1982). Acquisition of cognitive skill. *Psychological Review*, *89*, 369-406.

Anderson, J. R., Fincham, J. M., & Douglass, S. (1997). The role of examples and rules in the acquisition of a cognitive skill. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *23*, 932-945.

Anderson, J. R., Reder, L. M., & Simon, H. A. (2000). Applications and misapplications of cognitive psychology to mathematics education. *Texas Educational Review*, *1*, 29-49.

Atkins, J. (2005). The association between the use of Accelerated Math and students' math achievement. Doctoral thesis. East Tennesse State University. (http://etd-submit.etsu.edu/etd/theses/available/etd-0504105-130335/)

Atkinson, R. C. & Fletcher, J. (1972). Teaching children to read with a computer. *Reading Teacher*, *25*, 319-327.

Bangert-Drowns, R. L., Kulik, C. C., Kulik, J. A., & Morgan, M. T. (1991). The instructional effect of feedback in test-like events. *Review of Educational Research*, *61*, 213-238.

Baroody, A. J. (2003). The development of adaptive expertise and flexibility: The integration of conceptual and procedural knowledge. In A. J. Baroody & A. Dowker (Ed.), *The development of arithmetic concepts and skills: Constructing adaptive expertise* (pp. 1-33). Mahwah, NJ: Erlbaum.

Bassoe, C. F. (1995). Automated diagnoses from clinical narratives: A medical system based on computerized medical records, natural language processing, and neural network technology. *Neural Networks*, *8*, 313-319.

Baumert, J., Schmitz, B., Roeder, P., & Sang, F. (1989). Zur Optimierung von Leistungsforderung und Chancenausgleich in Schulklassen: Explorative Untersuchungen mittels HYPAG. / Maximizing achievement and equal opportunity in the classroom: Explorative studies with HYPAG. *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie*, *21*, 201-222.

Brown, J. S. & VanLehn, K. (1981). Repair theory: A generative theory of bugs in procedural skills. *Cognitive Science*, *4*, 379-426.

Brophy, J. (1986). Teacher influences on student achievement. *American Psychologist*, *41*, 1069-1077.

Carpenter, T. P., Fennema, E., Franke, M. L., Levi, L., & Empson, S. (1999). *Children's mathematics: Cognitively Guided Instruction*. Portmouth, NH: Heinemann.

Carstensen, C. & Rost, J. (2003). MULTIRA – a software system for multidimensional Rasch models. Kiel: IPN. (http://www.multira.de)

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Hillsdale, NJ: Erlbaum.

Desoete, A. & Roeyers, H. (2006). Metacognitive macroevaluations in mathematical problem solving. *Learning and Instruction*, *16*, 12-25.

Desoete, A., Roeyers, H., & De Clercq, A. (2003). Can offline metacognition enhance mathematical problem solving? *Journal of Educational Psychology, 95*, 188-200.

Elia, I., Gagatsis, A., & Demetriou, A. (2007). The effects of different modes of representation on the solution of one-step additive problems. *Learning and Instruction*, *17*, 658-672.

Forsstrom, J. J. & Dalton, K. J. (1995). Artificial neural networks for decision support in clinical medicine. *Annals of Medicine*, *27*, 509-517.

Fuchs, L. S., Fuchs, D., Prentice, K., Hamlett., C. L., Finelli R. & Courey, S. J. (2004). Enhancing mathematical problem solving among third-grade students with schema-based instruction. *Journal of Educational Psychology*, *96*, 635 -647.

Fuson, K. C, Wearne, D., Hiebert, J. C., Murray, H. G., Human, P. G., Olivier, A. I., Carpenter, T. P., & Fennema, E. (1997). Children's conceptual structures for mulitdigit numbers and methods of multidigit addition and subtraction. *Journal for Research in Mathematics Education*, *28*, 130-162.

Gagne, R. M. (1962). The acquisition of knowledge. *Psychological Review*, *69*, 355-365.

Gagnon & Maccini (2001). Preparing students with disabilities for algebra. *Teaching for Exceptional Children*, *34*, 8-15.

Greenwood, C. R. (1991). Longitudinal analysis of time, engagement, and achievement in at-risk versus non-risk students. *Exceptional Children*, *56*, 521-535.

Hecht, S. A., Torgeson, J. K., Wagner, R. K., & Rashotte, C. A. (2001). The relations between phonological processing abilities and emerging individual differences in mathematical computation skills. A longitudinal study from second to fifth grades. *Journal of Experimental Child Psychology*, *79*, 192-227.

Heirdsfield, A. M. & Cooper, T. J. (2002). Flexibility and inflexibility in accurate mental addition and subtraction: two case studies. *The Journal of Mathematical Behavior*, *21*, 57-74.

Helmke, A. (1988). Leistungssteigerung und Ausgleich von Leistungsunterschieden in Schulklassen: unvereinbare Ziele? *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie*, *20*, 45-76.

Hiebert, J. & Wearne, D. (1996). Instruction, understanding, and skill in multidigit addition and subtraction. *Cognition and Instruction*, *14*, 251-283.

Hilgard, E. R., Irvine, R. P., & Whipple, J. E. (1953). Rote memorization, understanding, and transfer: An extension of Katona's card-trick experiments. *Journal of Experimental Psychology*, *46*, 288-292.

Jamison, D., Suppes, P. &, Wells, S. (1974). The effectiveness of alternative instructional media: A survey. *Review of Educational Research*, *44*, 1-67.

Jitendra, A. K., Griffin, C. C., Haria, P., Leh,J., Adams, A., & Kaduvettoor, A. (2007). A comparison of single and multiple strategy instruction on third-grade students' mathematical problem solving. *Journal of Educational Psychology*, *99*, 115-127.

Kail, R. & Hall, L. K. (1999). Sources of developmental change in children's word-problem performance. *Journal of Educational Psychology*, *91*, 660-668.

Kirschner, P. A., Sweller, J., & Clark, R. E. (2006). Why minimal guidance during instruction does not work: An analysis of the failure of constructivist, discovery, problem-based, experiential, and inquiry-based teaching. *Educational psychologist*, *41*, 75-86.

Koedinger, K. R., Anderson, J. R., Hadley, W. H., & Mark, M. (1997). Intelligent tutoring goes to school in the big city. *International Journal of Artificial Intelligence in Education*, *8*, 30-43.

Kroesbergen, E. H. & Van Luit, J. E. H. (2003). Mathematics interventions for children with special educational needs: A Meta-Analysis. *Remedial and special education*, *24*, 97-114.

Ku, H.-Y., Harter, C. A., Liu, P.-L., Thompson, L., & Cheng, Y.-C. (2007). The effects of individually personalized computer-based instructional program on solving mathematics problems. *Computers in Human Behavior*, *23*, 1195-1210.

Kulik, J. A. (1994). Meta-analytic studies of findings on computer-based instruction. In E. L. Baker & H. F. O'Neil (Eds.), *Technology assessment in education and training* (pp. 9-33). Hillsdale, NJ: Erlbaum.

Laue, C. & Putz-Osterloh, W. (2002). Computergestütztes Lernen in Mathematik bei Grundschülern. In Birgit Spinath & Elke Heise (Hg.) *Pädagogische Psychologie unter gewandelten gesellschaftlichen Bedingungen* (pp. 68-83). Hamburg: Verlag Dr. Kovac.

Lehmann, R. & Seeber, S. (2005). *"Accelerated Mathematics" in grades 4 through 6*. http://zope.ebf.hu-berlin.de/document/

Lovett, M. C. & Anderson, J. R. (1996). History of success and current context in problem solving: Combined influences on operator selection. *Cognitive Psychology*, *31*, 168-217.

Luchins, A. S. (1942). Mechanization in problem solving: The effect of einstellung. *Psychological Monographs*, *54*, 95-111.

Mevarech, Z. R. & Rich, Y. (1985). Effects of computer-assisted mathematics instruction on disadvantaged pupil's cognitive and affective development. *Journal of Educational Research*, *79*, 5-11.

National Council of Teachers of Mathematics (2000). *Principles and standards for school mathematics: Standards 2000*. Reston, VA: Author.

Paas, F. G. W. C. & Merriënboer, J. J. G. (1994). Variability of worked examples and transfer of geometrical problem-solving skills: A cognitive-load approach. *Journal of Educational Psychology*, *86*, 122-133.

Pearl, J. (2001). *Probabilistic reasoning in intelligent systems*. San Francisco, CA: Morgan Kaufmann.

Rohrer, D. & Taylor, K. (2007). The shuffling of mathematics problems improves learning. *Instructional Science*, *35*, 481-498.

Roick, T., Gölitz, D., & Hasselhorn, M. (2004). *DEMAT 3+. Deutscher Mathematiktest für dritte Klassen*. Göttingen: Beltz Test.

Rost, J. (2004). *Lehrbuch Testtheorie – Testkonstruktion*. Bern: Huber.

Schoppek, W. (2006). A hierarchy of elementary arithmetic skills as a basis for individualised instruction. (http://www.uni-bayreuth.de/departments/psychologie/schoppek/arithskill.pdf)

Sherin, B. & Fuson, K. C. (2005). Multiplication strategies and the appropriation of computational resources. *Journal for Research in Mathematics Education*, *36*, 347-395.

Star, J. R. & Rittle-Johnson, B. (2008). Flexibility in problem solving: The case of equation solving. *Learning and Instruction*, *18*, 565-579.

Stark, R., Mandl, H., Gruber, H., & Renkl, A. (1999). Instructional means to overcome transfer problems in the domain of economics: Empirical studies. *International Journal of Educational Research*, *31*, 591-609.

Sweller, J. (1988). Cognitive load during problem solving: Effects on learning. *Cognitive Science*, *12*, 257-285.

Treiber, B., Weinert, F. E., & Groeben, N. (1982). Unterrichtsqualität, Leistungsniveau von Schulklassen und individueller Lernfortschritt. *Zeitschrift für Pädagogik*, *27*, 65-75.

Tronsky, L. N. (2005). Strategy use, the development of automaticity, and working memory involvement in complex multiplication. *Memory & Cognition*, *33*, 927-940.

Van Essen, G. & Hamaker, C. (1990). Using self-generated drawings to solve arithmetic word problems. *The Journal of Educational Research, 83,* 301 -312.

Verschaffel, L., DeCorte, E., & Vierstraete, H. (1999). Upper elementary school pupils' difficulties in modeling and solving nonstandard additive word problems involving ordinal numbers. *Journal for Research in Mathematics Education*, *30*, 265-285.

White, Richard T. (1976). Effects of guidance, sequence, and attribute-treatment interactions on learning, retention, and transfer of hierarchically ordered skills. *Instructional Science*, *5*, 133-152.

Ysseldyke, J., Spicuzza, R., Kosciolek, S., & Boys, C. (2003). Effects of a learning information system on mathematics achievement and classroom structure. *The Journal of Educational Research*, *96*, 163-173.