Instr Sci
DOI 10.1007/s11251-015-9350-0

Using mental computation training to improve complex mathematical performance

Allison S. Liu[1,2] · Arava Y. Kallai[3] · Christian D. Schunn[1,2] · Julie A. Fiez[1,2,4]

Received: 16 August 2014 / Accepted: 13 April 2015

Abstract Mathematical fluency is important for academic and mathematical success. Fluency training programs have typically focused on fostering retrieval, which leads to math performance that does not reliably transfer to non-trained problems. More recent studies have focused on training number understanding and representational precision, but few have directly investigated whether training improvements also transfer to more advanced mathematics. In one previous study, university undergraduates who extensively trained on mental computation demonstrated improvements on a complex mathematics test. These improvements were also associated with changes in number representation precision. Because such far transfer is both rare and educationally important, we investigated whether these transfer and precision effects would occur when using a more diverse population and after removing several features of the mental computation training that are difficult to implement in classrooms. Trained participants showed significant, robust improvements, suggesting that mental computation training can reliably lead to mathematical transfer and improvements in number representation precision.

Keywords Mathematical fluency · Mental computation · Number representation · Number understanding · Transfer

✉ Allison S. Liu
asl36@pitt.edu

[1] Department of Psychology, University of Pittsburgh, Pittsburgh, PA, USA

[2] Learning Research and Development Center 823, University of Pittsburgh, 3939 O'Hara Street, Pittsburgh, PA 15260, USA

[3] Department of Psychology, Ben-Gurion University of the Negev, Beer-Sheva, Israel

[4] Department of Neuroscience, University of Pittsburgh, Pittsburgh, PA, USA

Published online: 21 April 2015

Introduction

Mathematical fluency is important for mathematical competence and future academic success (Mazzocco et al. 2008; Pellegrino and Goldman 1987; Resnick 1983). As a result, fluency, typically defined as the ability to perform mathematics quickly, accurately, and flexibly (National Council of Teachers of Mathematics 2000; National Governors Association Center for Best Practices & Council of Chief State School Officers 2010), has been the focus of many mathematical training programs that encourage the memorization and fast retrieval of basic math facts (i.e., single-digit addition and multiplication). These retrieval-based training programs have been successful in fostering quick and accurate fluency. For example, elementary and middle-school students who used the training program FASST Math gained more addition and multiplication math facts that could be retrieved in less than one second than students who experienced only their regular mathematics curriculum (Scholastic Inc. 2005). Providing converging evidence at the mechanistic level, individuals with higher mathematical problem solving competencies showed greater activation levels in brain regions associated with retrieval processes during calculation tasks (Price et al. 2013). The brain regions activated during retrieval are also activated during transfer between arithmetic operations (Ischebeck et al. 2009). However, improvements from retrieval training have been shown to apply only to trained, or very similar, problems, suggesting that such training programs do not always lead to flexible mathematical fluency (Bajicet al. 2011; Imbo and Vandierendonck 2008; Rickard et al. 1994).

The primary benefits of retrieval training appear to lie in the reduction of working memory load rather than increased numerical understanding (Pellegrino and Goldman 1987; Resnick 1983; Sweller et al. 1983; Woodward 2006). After training, brain regions involved during basic mathematical calculations shift away from fronto-parietal networks that are involved in attentional processing to the left angular gyrus, a region that has been associated with retrieval processes (Delazer et al. 2003; Ischebeck et al. 2007). While a lower working memory load can allow attentional resources to be allocated toward more complex math strategies (Imbo et al. 2007), load reduction alone may not be sufficient for transfer if conceptual knowledge is needed to apply the trained procedural knowledge to novel numerical situations. It is possible that a training program that encourages deeper numerical knowledge would foster mathematical fluency that is flexible, as well as quick and accurate.

In a recent study, Kallai et al. (2011) found that far transfer to untrained, more complex mathematics is possible after mathematical fluency training that encourages deeper numerical knowledge. Participants were trained through mental computation [i.e., the process of performing arithmetic operations without external devices (Sowder 1988)], a method thought to build flexible, transferable knowledge about numerical symbols and the quantities they represent (Markovits and Sowder 1994; Reys 1984; Sowder 1992; Thompson 1999). Participants solved multi-digit addition and subtraction problems over several training sessions. Particular problems rarely repeated, such that participants could not simply memorize answers (Schunn et al. 1997). To encourage participants to utilize whole number quantity processing rather than rote, numerically-meaningless calculations, the problems were horizontally formatted, and participants had a short amount of time to solve the problems. The training also included immediate feedback, high uncertainty, and rewards, which have been shown to modulate a basal ganglia learning system that is involved in motor and cognitive skill learning and representational change (see Tricomi and Fiez 2008 for a review). It is not surprising that the training group greatly improved on the training task itself. However, the training group also showed improvements on mathematical tasks that were very different from the training task: the mental computation group

demonstrated significant improvements on a complex mathematics test involving reasoning about ratios/proportions, algebraic equations, and probabilities. Thus, improvements from the mental computation training transferred to problems that required more advanced mathematics than the addition and subtraction that was trained.

The study included important methodological controls to rule out trivial explanations for the far transfer results. Improvements in the basic training task and the complex mathematics task were not associated with a simple increase in component (i.e., single-digit) math facts retrieval efficiency. Two alternative forms of the complex mathematics test were given in counter-balanced order across participants as pre- and post-tests. The training group showed greater gains than a test–retest group who completed no training. Thus, simple practice effects with complex mathematics tests were ruled out as the basis of gains in the training group.

Additionally, the study included an active control group that was trained to select one of two visually-presented, double-digit numbers based on a colored symbol cue, and to type that number (a numerically-meaningless task that controlled for exposure to and typing of numerical stimuli). The group was trained over the same time course as the mental computation group and received an equivalent monetary reward for the task. Again, the training group showed greater gains in complex mathematics than the active control group, ruling out changing motivation levels during the test or covert practice on complex mathematics between training sessions. This control group also ruled out simple raw exposure to double-digit numbers as the basis of transfer.

What might be the basis of this far transfer? We argue that it involves changing representations of underlying quantities (Kallai et al. 2011). Humans' numerical representations are thought to be analogous to a mental number line (Dehaene et al. 1993; Gallistel and Gelman 1992). Because these representations are approximate, representations of close quantities on the number line overlap, making close quantities harder to discriminate than distant quantities (Dehaene et al. 1990; Moyer and Landauer 1967). In the brain, the horizontal segment of the intraparietal sulcus appears to be where quantity representations are localized; this region is activated during number manipulation, with greater activation when tasks require more quantity processing (Dehaene et al. 2003). People rely on these quantity representations during symbolic numerical tasks when the task requires processing of the numerical meaning of symbols (Dehaene and Marques 2002; Gallistel and Gelman 2005). Furthermore, greater precision in these representations has been associated with higher symbolic math performance in both children and adults (e.g., Halberda et al. 2012; Halberda et al. 2008; Holloway and Ansari 2009; Lourenco et al. 2012) and with later math achievement (e.g., Gilmore et al. 2010), making improvements in precision seem a strong candidate for improving mathematical fluency. Several studies have successfully trained quantity representations in children and adults (e.g., Fischer et al. 2011; Obersteiner et al. 2013; Park and Brannon 2013; Ramani and Siegler 2011; Whyte and Bull 2008), leading to significant mathematical improvements. With regard to flexibility, however, most training studies have investigated transfer to more basic arithmetic skills, such as single- or double-digit addition and subtraction, and most studies investigating the relationship between precision and higher-order mathematics tend to be correlational in nature. In contrast, the Kallai et al. (2011) study demonstrated pre-post changes on behavioral and neuroimaging measures of precision in the approximate representation of multi-digit numbers (greater than the test–retest and the active control groups), and found that this change in numerical precision was associated with the degree of improvement in complex mathematics.

Such instances of far transfer are educationally important and rare, and invite important questions of robustness across populations and training variations. While the Kallai et al. (2011) study demonstrated fast, accurate, and flexible math ability, it was limited to a select

population of college students in a controlled laboratory situation. It is unknown whether the mental computation training used in this study would continue to be effective in improving math performance and transfer to complex math skills in more realistic arithmetic-training situations, which may involve a more varied participant pool or variations of the original training paradigm that might be more easily implemented in classrooms. In the current report, we tested the robustness of these training and transfer effects across two studies. In one study, we recruited a heterogeneous group of participants through Amazon Mechanical Turk (AMT) and compared participants who completed one of two different Internet-adapted variations of our mental computation training to participants who completed no training. In the second study, which was lab-based, we compared participants who completed the original version of the mental computation training to participants who completed one of three training variations, each of which eliminated one core learning feature from the original training (i.e., immediate feedback, high uncertainty, or rewards for accurate performance). In both studies, we tested participants' performance on the complex math test and their symbolic representational precision before and after training. It should be noted that our studies were not exact replications of Kallai et al.'s (2011) previous study. Because we were primarily interested in testing the robustness of the previous effects under various conditions, it was important to make changes to the previous experiment's design. To foreshadow our results, the AMT training groups and all three variations of the training task showed greater performance on the complex math test after training and improvements in symbolic precision, suggesting that the mental computation training can robustly produce important mathematical improvements.

Study 1—generalizing populations

Materials & methods

Participants

Seventy-one adults participated in the study. Participants were recruited through AMT, an online, crowd-sourced participant pool. Recruiters are able to post tasks on the AMT website, and workers choose tasks to complete. If a worker satisfactorily completes a task, then recruiters grant approval and payment to that worker. All AMT participants recruited in study 1 were required to be located in the U.S. and to have an approval rate of 95 % or higher, which has been found to ensure high-quality AMT data (Peer et al. 2014). In general, studies conducted using AMT have consistently replicated results from the lab, but have allowed for better generalization to broader ages and demographics than are typically found in university-based lab studies (Buhrmester et al. 2011; Goodman et al. 2013; Mason and Suri 2012; Paolacci et al. 2010; Simcox and Fiez 2013). Thirty-seven participants completed the training conditions for $21 base pay with a performance bonus of $1–$7. These participants were randomly assigned to one of two training conditions (described below): 20 to Window training (8 female), and 17 to Ratio training (7 female). In addition, 34 participants (11 female) completed only the pre- and post-tests (No-Training condition) for $4 base pay with a performance bonus of $0.06–$0.60. Our No-Training condition completed only the pre- and post-tests because Kallai et al.'s (2011) had already compared the mental computation training against an active control group, and that study had ruled out simple exposure with double-digit numbers and motivational differences as bases of the observed transfer effects. Studies in other cognitive domains also show significant test–

retest effects (Salthouse 2010), and this older population may have had less prior or recent exposure to complex mathematics, suggesting test–retest to be an important control condition to include in this study.

All participants reported being native english speakers. Participants were restricted to non-quantitative prior or current college majors (e.g., engineering, science) to avoid ceiling effects. Reported education levels for each condition ranged from high-school/GED level to master's degrees (Window: 4 HS/GED, 6 associates/bachelor's-in-progress, 10 bachelor's degrees; Ratio: 6 HS/GED, 1 associates/bachelor's-in-progress, 5 bachelor's degrees, 4 master's degrees, 1 unreported; No-Training: 4 HS/GED, 9 associates/bachelor's-in-progress, 9 bachelor's degrees, 1 master's degree, 11 unreported). Ages of the participants were higher overall and more varied than in Kallai et al.'s (2011) study (Kallai, et al. : $M = 20.9$, $SD = 1.6$ years; Window: $M = 30.1$, $SD = 8.6$ years; Ratio: $M = 30.3$, $SD = 9.5$ years; No-Training: $M = 37.2$, $SD = 11.5$ years). The statistically significant age disparity between the two training conditions and the No-Training condition [$F(2, 58) = 3.53$, $p = 0.036$] was controlled for in our analyses. All participants gave informed consent before participating in the study.

Procedure

All tasks and training sessions in the study took place online. Participants in the No-Training condition completed only a pre-test session, and then a post-test session 6 days later. Participants in the two training conditions completed seven 1-h sessions. Each session was separated by a minimum of 8 h to encourage benefits of consolidation during sleep and a maximum of 48 h to ensure steady progress. Participants completed a pre-test session, followed by five sessions of training (either window or ratio training), followed by a post-test session. For all conditions, the pre-test session included the complex math, arithmetic fluency (as a control measure), and number comparison (to measure representation precision) tasks, and the post-test session included the complex math and number comparison tasks.

Tasks

Estimation training task During each training session, participants completed five sets of double-digit addition problems, then five similarly-designed sets of double-digit subtraction problems, with 40 problems per set. Operands in the problem were shown sequentially to encourage holistic number representation: the first operand was shown for 0.5 s, followed by an operation symbol (+ or −) for 0.25 s, followed by the second operand for 0.5 s. Participants were then shown a blank screen and had 2 s to type in their response using the number keypad on the keyboard (see Fig. 1). In the first training session, the first set in each operation was a warm-up set that only included single-digit problems. Feedback was given in the form of green checks for correct responses (with more checks indicating a response closer to the exact answer) and a red "X" for incorrect responses. Participants also saw their total percent of correct answers, mean response time, and number of points they received at the end of each set.

If participants achieved 90 % accuracy on a set, the difficulty level (altered by manipulating the range of the acceptable responses) increased during the next set of problems, up to a difficulty level of 5. We explored two methods for adaptive training. In the Window training condition, participants were required to answer within five of the exact answer at difficulty level 1 (e.g., if the exact solution was "100", then any response from 95 to 105 would be correct), down to an estimation window of one at difficulty 5 (e.g., if the exact solution was "15", then any response from 14 to 16 would be correct). In the Ratio

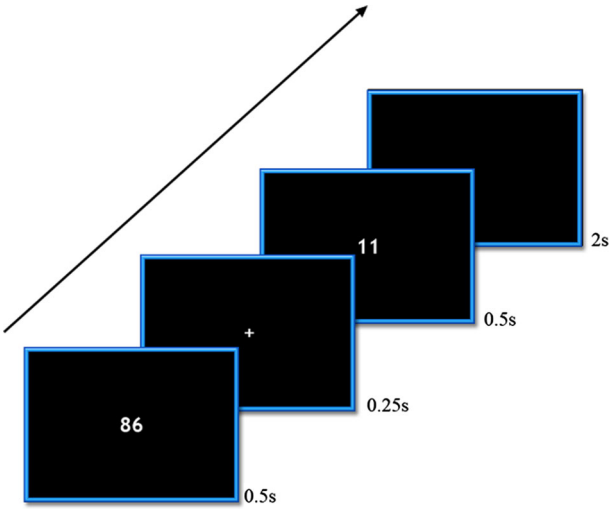

Fig. 1 Schematic of training task trials

condition, participants were required to answer within 8 % (rounded to the nearest whole number) of the exact answer at difficulty level 1 (e.g., if the exact solution was "100", then any response from 92 to 108 would be correct); with each difficulty level, the ratio decreased by 2 %, such that the exact response was required at difficulty level 5. This modified estimation-based approach to Kallai and et al.'s (2011) training was used to adapt to any heterogeneity in participants' starting arithmetic knowledge, allowing all participants to experience difficult training (thereby maximizing learning effects) while still staying within the targeted double-digit operands.

Arithmetic fluency task As a control measure of participants' traditional multi-digit arithmetic skill before training, participants solved multi-digit addition and subtraction problems for their exact solutions as quickly and as accurately as possible on a computer during the pre-test session. Four sets of problems were used, with 16 problems per set. The first two sets of problems involved addition and subtraction, respectively, with one double-digit operand and one single-digit operand. The last two sets of problems involved addition and subtraction, respectively, with two double-digit operands. Both operands were shown simultaneously at the center of the screen for 10 s. Participants were required to type their responses and press the Enter key to register their response before the problem disappeared from the screen. Participants' percent accuracy was shown after each set. Accuracy and response times (the time of Enter key-press) were collected for each problem.

Complex math task To measure the effect of training on complex mathematical skills, this computerized task consisted of 16 multiple-choice mathematical problems taken from a Scholastic Assessment Test (SAT) preparation book, the same assessment used in the Kallai et al. (2011) study. The SAT is a widely used college entrance exam in the U.S., and represents high-level mathematical reasoning competencies of broad importance in college coursework.

The problems involved complex mathematical computations and conceptual reasoning, such as algebraic computation and statistical reasoning, in contrast to the procedural multi-digit addition and subtraction skills practiced during training. Isomorphs of the original 16

problems were created and used as a second version of the test. Each participant completed one version of the test during the pretest session, and the other version of the test during the post-test session, with order of versions counterbalanced across participants. Participants had 30 min to solve the problems as quickly and as accurately as possible. Four questions were shown per page, and participants could not skip back and forth between pages. Only accuracy scores were collected from the test, as we could not control environmental factors that could potentially confound response time measures for longer duration problems in the online environment. Two example problems are as follows. These particular examples showed training improvements and highlight how multi-digit addition or subtraction is not a salient element of the complex mathematics test:

- Of Team X's victories this year, 80 % were at home. If Team X has won a total of 30 games this year, how many of those games were won away from home? (A) 4 (B) 5 (C) 6 (D) 24 (E) 26 (F) None of the above
- There are n players on a team. If, among those players, p % have no helmet, which of the following general expressions represents the number of players who have a helmet? (A) np (B) $.01np$ (C) $(100 - p)n/100$ (D) $(1 - p)n/0.01$ (E) $100(1 - p)n$ (F) None of the above

Number Comparison task A number comparison task was given pre and post to assess representational change due to training (see Dehaene 1992). Two numbers were presented sequentially on a computer screen. Participants were instructed to indicate whether the second number was smaller or larger than the first number (called the "standard") by pressing the "S" or "L" key, respectively, on the keyboard. The standard was presented in the instructions at the start of each block and did not change within the block. Participants completed nine blocks of problems, with 16 trials per block. The first block was a practice block, which used the standard of "41": and provided trial-by-trial feedback. The next eight blocks consisted of randomly ordered blocks using the standards of 18, 25, 32, and 49 (with 2 blocks per standard). Percent accuracy was provided at the end of each block. Each standard number was paired with 16 comparison numbers; half of the comparison numbers were smaller than the standard, and half were larger. The comparison numbers for each standard were created by multiplying the standard number by one of sixteen ratios (0.6, 0.687, 0.718, 0.781, 0.812, 0.875, 0.906, 0.968, 1.031, 1.125, 1.156, 1.218, 1.25, 1.333, 1.406, and 1.437), and then rounding to the nearest natural number. For each trial, the standard was presented for 0.4 s, followed by a fixation cue ("#") for 0.3 s, followed by the comparison number for 0.7 s, followed by an additional response period of 1.3 s (during which participants responded with the "S" or "L" key). Although environmental factors may have influenced participants' response times during this task as well, it was possible to detect and control for such influences as the task involved relatively brief response times (i.e., if participants inputted no response to many consecutive trials). Thus, both accuracy and response times were collected.

Results

Training results

To verify that the training conditions improved in the trained double-digit arithmetic, we compared participants' mean accuracy on the first block of the first training session and

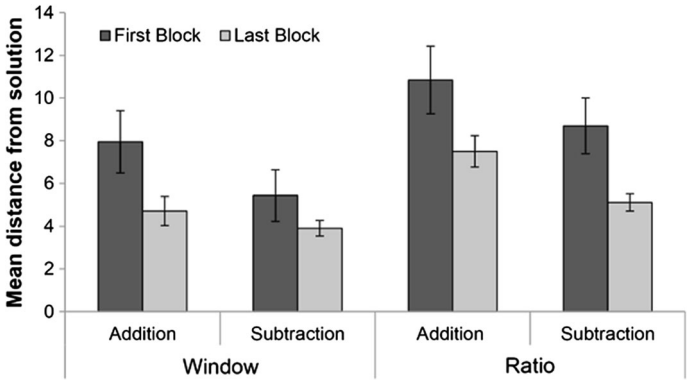

Fig. 2 Mean distance from exact solution (and within SE bars) at first and last block of training, separated by training condition (Window, Ratio) and operation

accuracy on the last block of the last training session. Because participants were trained to estimate the solution to problems, accuracy was measured as the distance from the exact problem solution; a lower distance score means that a participant was more accurate in their answers. Analyses were computed separately for addition problems and subtraction problems.

Both training conditions began at similar levels of accuracy for addition [$F(1, 35) = 1.81$, $p = 0.19$] and subtraction [$F(1, 35) = 3.35$, $p = 0.08$]. A Training Condition (window, ratio) $\times$ Block (first, last) repeated-measures ANOVA yielded a significant main effect of Block for both addition [$F(1, 35) = 13.96$, $p = 0.001$, $d = 0.63$] and subtraction problems [$F(1, 35) = 9.15$, $p = 0.005$, $d = 0.60$], such that accuracy on the last block of each operation's problems was better than accuracy on the first block of problems. For subtraction problems, a main effect of Group showed that the Window condition performed better than the Ratio condition across the two blocks [$F(1, 35) = 4.90$, $p = 0.033$]. However, there was no interaction between Training Condition and Block for either operation [addition: $F < 1$; subtraction: $F(1, 35) = 1.47$, $p = 0.23$, $\eta_p^2 = 0.04$], suggesting that both training conditions equally improved multi-digit addition and subtraction accuracy (see Fig. 2).

Complex math

Based on pre-test accuracy, the two complex math test versions differed in their difficulty. Therefore, we analyzed accuracy performance using standardized scores based on the given form, using the mean and standard deviation of the pre-tests. There were no significant differences between the three conditions (Window, Ratio, No-Training) in pre-test accuracy [$F < 1$]. There were also no significant differences between the two training conditions in either pre-test [$F < 1$] or post-test accuracy [$F(1, 35) = 1.82$, $p = 0.19$]. The Window and Ratio conditions correctly answered approximately 7 and 6 questions (as a mean out of 16), respectively, at pre-test, while the No-Training condition correctly answered approximately 7 questions. Because the two training conditions did not differ in either training accuracy gains or Complex Math pre- and post-test scores, the two training conditions were combined for increased statistical power (and labeled as the Training condition) in the remainder of the analyses.

The Kallai et al. (2011) study used completion time as a covariate in pre-post Complex Math accuracy analyses, but this variable was not available in the current study. Instead, we used other available individual difference covariates. We elected to run a one-way ANCOVA instead of a repeated-measures ANCOVA for two reasons. Participants were likely to start with different levels of arithmetic ability (rather than differing just on complex mathematics skills), and comparing simple changes from pre-test to post-test ignores this important source of variance. Furthermore, ANCOVA designs tend to have greater statistical power than repeated measures analyses for measuring the effects of interventions (Delaney and Maxwell 1981; Dimitrov and Rumrill 2003; Van Breukelen 2006). We ran the one-way ANCOVA on post-test Complex Math accuracy, with Condition (Training, No-Training) as a between-subjects factor, and pre-test Arithmetic Fluency accuracy, pre-test Complex Math accuracy, and age as covariates to control for participants' prior basic and complex mathematical ability and for the age differences between the groups (the same results hold whether or not age is included in the analysis). The Training condition ($M = 0.36$, $SD = 1.03$) scored significantly higher on the post-test compared to the No-Training condition ($M = -0.04$, $SD = 1.00$) [$F(1, 54) = 4.76$, $p = 0.03$, $d = 0.39$] (see Fig. 3): the Window and Ratio conditions correctly answered an average of 9 and 7 questions (out of 16), respectively, while the No-Training condition correctly answered an average of 7 questions (i.e., showed no gains, as was found previously, from simple test–retest effects). To compare our results more directly to Kallai et al.'s (2011) study, we also ran a Condition (Training, No Training) X Session (pre, post) one-tailed repeated measures ANCOVA with Condition as a between-subjects factor, Session as a within-subjects factor, and pre-test arithmetic fluency accuracy as a covariate to control for basic arithmetic ability. We used a one-tailed test because we were specifically interested in whether the training condition would outperform the control condition, as was found by Kallai et al.'s. The repeated-measures ANCOVA showed a significant interaction between Condition and Session [$F(1, 68) = 2.86$, $p = 0.048$, $\eta_p^2 = 0.04$], such that the Training condition improved significantly more than the No-Training condition from pre-test to post-test. There was no main effect of Condition [$F(1, 68) = 2.06$, $p = 0.08$, $d = 0.35$] or Session [$F(1, 68) = 0.14$, $p = 0.36$, $d = -0.25$]. The interaction also holds when age is added as a covariate [$F(1, 57) = 3.16$, $p = 0.041$, $\eta_p^2 = 0.05$]. Thus, the training on multi-digit addition and subtraction appeared to improve

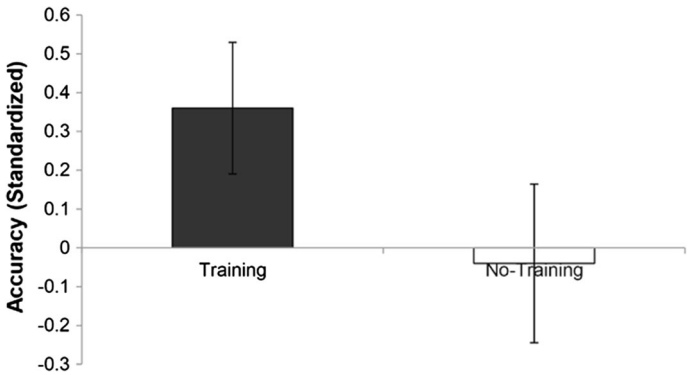

Fig. 3 Adjusted mean standardized complex math post-test accuracy (and SE bars) for training (Window, Ratio) and No-Training conditions

participants' complex mathematical ability, above any improvements that may have been caused by familiarity due to pre-test exposure to problems of this type.

Number comparison

To assess changes in numerical representation precision, a Condition (Training, No-Training) X Session (pre, post) X Distance (near, far) repeated-measures ANOVA was run on both accuracy and response times on the number comparison task, with condition as a between-subjects factor, and session and distance as within-subject factors. We separated the ratios used in the task into two levels, based on their distance from the 1:1 ratio: near (ratios of 0.781, 0.812, 0.875, 0.906, 0.968, 1.031, 1.125, 1.156, 1.218, and 1.25) and far (ratios of 0.6, 0.687, 0.718, 1.333, 1.406, and 1.437). The divide was chosen based on prior data; pre-training accuracy and response times showed strong differentiation for quantities beyond ratios of 0.781 and 1.25. Near versus far task performance difference was used as our primary ANS acuity measure, based on recent findings that such a measure is a more reliable measure of ANS acuity than the more commonly used Weber fraction (Inglis and Gilmore 2014). There were no significant differences between the two training conditions in either pre-test [$F < 1$ for both near and far ratios] or post-test accuracy [$F < 1$ for both near and far ratios], or in pre-test [near: $F(1, 35) = 2.17$, $p = 0.15$; far: $F(1, 35) = 2.45$, $p = 0.13$] or post-test response times [near: $F(1, 35) = 2.16$, $p = 0.15$; far: $F(1, 35) = 2.17$, $p = 0.15$]; thus, the two training conditions were combined (and labeled as the training condition) as was done in the complex math analyses.

There were significant main effects of session [$F(1, 69) = 6.84$, $p = 0.01$, $\eta_p^2 = 0.09$] and Distance [$F(1, 69) = 67.5$, $p < 0.001$, $\eta_p^2 = 0.49$] for task accuracy: accuracy was higher at post-test ($M = 0.97$; $SD = 0.03$) than pre-test ($M = 0.96$; $SD = 0.03$), and higher for far ratios ($M = 0.98$; $SD = 0.03$) than near ratios ($M = 0.95$; $SD = 0.03$). There was no three-way interaction of Group X Session X Distance [$F(1, 69) = 1.85$, $p = 0.18$, $\eta_p^2 = 0.03$]; participants simply became more accurate with second exposure to the task regardless of training. However, the response time measures showed the predicted and theoretically critical three-way interaction [$F(1, 69) = 4.59$, $p = 0.036$, $\eta_p^2 = 0.70$]. An analysis of simple effects showed that the Session X Distance interaction was significant for the Training group [$F(1, 36) = 4.27$, $p = 0.046$, $\eta_p^2 = 0.11$], but not for the No-Training group [$F(1, 33) = 1.09$, $p = 0.30$, $\eta_p^2 = 0.03$]. The results of the analyses were the same when prior basic and complex mathematical ability (i.e., pre-test Arithmetic Fluency accuracy and pre-test Complex Math accuracy) were controlled. The response time for near ratios significantly decreased from pre-test to post-test in the Training condition only (see Fig. 4), suggesting that training made number representations more precise, allowing participants to more quickly distinguish close quantities.

Discussion

As predicted, participants who completed the mental computation training significantly improved on the complex math test compared to participants who did not complete any training. The training group could also distinguish close symbolic numbers more quickly after training, suggesting that the training improved their symbolic representational precision. Our results build on those from the Kallai et al. study (2011), providing an important second demonstration of a rare far transfer result. Furthermore, the far transfer

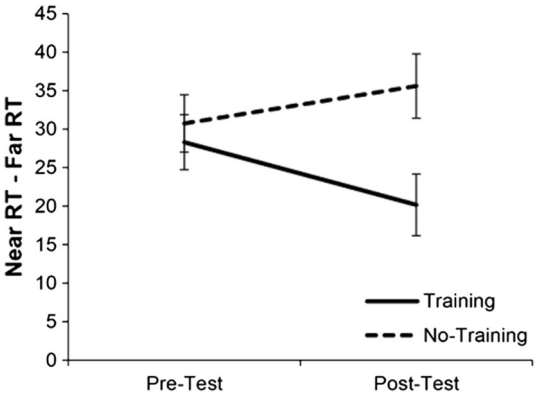

Fig. 4 Mean difference (and within SE bars) between near and far ratio response times on number comparison pre-test and post-test for training and No-Training groups. Values closer to 0 indicate a smaller difference between far and near ratios

result was obtained in a population with a much wider range of backgrounds, suggesting generalizability of these effects to a broad range of training populations.

Still, the tested mental computation training relies on a specific approach that may not be pragmatic in all learning contexts. This specific training approach was designed using basic research on effective learning environment characteristics, specifically to promote learning and representational change: (1) use of immediate feedback (Tricomi et al. 2004, (2) use of high uncertainty (Aron et al. 2004; Berns et al. 2001, (3) use of monetary rewards for accurate performance (Delgado et al. 2004). These characteristics may not be available in typical educational environments. In addition, the Training condition received a higher payment than the No-Training condition ($21 versus $4 base payment) in study 1, which may have motivated the Training condition participants and accounted for their higher performance. To determine whether the core training approach would retain its effectiveness even without these key features of feedback, uncertainty, or monetary rewards, we tested three variations of the original training in study 2. The study also serves as another demonstration of the representational change and far transfer results, further ruling out statistical flukes as an explanation of such rarely-obtained far transfer results.

Study 2—robustness across training characteristics

Materials & methods

Participants

Eighty college students or recent graduates participated in one of four training variations [Base-Training, No-Feedback, High-Certainty, and No-Reward (explained in detail below) with 20 students (10 females) per condition] for a base payment of $115, with additional training and pre-/post-test performance bonuses for a mean total pay of $205 (with the exception of the No-Reward training condition, who received a base payment equal to $205 and additional pre-/post-test performance bonuses). An a priori power analysis was conducted using the data from study 1, to ensure that we could detect any significant differences between the four training variations. With an alpha of 0.05, a power of 0.95, and an effect size of $\eta_p^2 = 0.040$, the sample size needed is approximately 76 participants for a repeated-measures ANOVA. Thus, our sample size of 80 participants should be

sufficient to test whether each of the core training features is necessary to obtain the training effect.

Matching the demographics of Kallai et al.'s (2011), all participants were right-handed and native english speakers between the ages of 18–25 years of age. To avoid ceiling effects, participants were screened to be non-experts in math, defined as having a non-quantitative major and having a self-reported quantitative SAT score that ranked them within the 75–93 percentile of college-bound seniors. The four training conditions were matched for gender (10 females, 10 males) and for quantitative SAT score (mean scores for each group were between the 83rd–85th percentiles). Average ages were also similar across training conditions (Base-Training: $M = 20.6$ years, $SD = 1.3$ years; No-Feedback: $M = 20.7$, $SD = 1.4$ years; No-Reward: $M = 20.6$, $SD = 1.2$ years; High-Certainty: $M = 20.2$, $SD = 1.3$ years). All participants gave informed, written consent before participating in the study.

Procedure

Participants completed seven sessions over 9 days in the laboratory. Sessions 1 and 7 consisted of behavioral pre-tests and post-tests, both of which included the complex math and number comparison tasks. During sessions 2 through 6, participants spent approximately 30 min solving addition problems and 30 min solving subtraction problems.

Tasks

Exact training task The Base-Training condition's training task was similar to that used in study 1. Because study 2's participants were more homogeneous and relatively high in their starting arithmetic knowledge (as shown by their matched quantitative SAT scores), we asked participants to solve for exact solutions instead of the estimated solutions used in study 1. This more closely approximated the training task used in the Kallai et al. study (2011), and likely better matched typical classroom tasks.

Fifty problems were included in each problem set. Both multi-digit operands were displayed simultaneously in horizontal format to encourage holistic representations, followed by a response period during which participants typed their solution to the problem. Feedback was provided after the response window (three green checks after a correct response, three red crosses after an incorrect response, three white dashes after no response). A monetary bonus was also given for each correct response, with higher difficulty questions earning higher rewards. All participants completed a warm-up block, made up of single-digit problems, to familiarize them with the task.

Three variations of the Base-Training condition tested, in isolation, the importance of three core features of the previously tested learning environment (immediate feedback, high uncertainty, and rewards for correct performance), each of which is somewhat uncommon in typical educational learning environments. Pilot testing was used to insure the variations primarily varied the targeted dimension (e.g., only feedback rather than both feedback and certainty). The three training variations were:

- The No-Feedback condition: participants received no immediate trial-by-trial feedback, and instead were shown three blue squares after entering their responses, regardless of whether their responses were correct or incorrect. They were told their percentage of correct answers after every set of problems to make sure motivation was not also changed.

Table 1 Presentation times and earnings used for each difficulty level in study 2 (with high certainty condition parameters in parentheses)

	Digits	Stimulus (s)	Response (s)	Feedback (s)	ITI (s)	Earnings
Warm-up	Single/single	0.5	2	0.5	0.5	$0.01
Level 1	Double/single	0.7 (2.9)	2.2 (10)	0.5	0.5	$0.02 ($0.01)
Level 2	Double/double	0.9 (3.5)	2.6 (10)	0.5	0.5	$0.04 ($0.02)

- The High-Certainty condition: participants were exposed to each problem for 2.9–3.5 s depending on problem difficulty, and they were given an extended window of 10 s to enter their responses to each problem. Because participants in this condition were given approximately four times as much time as participants in the other conditions to view and solve the training problems, they were expected to achieve higher accuracy levels than participants in the other three training conditions. For this reason, they received a smaller monetary completion bonus, as they had a greater opportunity to earn a higher performance bonus. (see Table 1)
- The No-Reward condition: participants did not receive a monetary bonus for correct responses during the training task (when rewards should have the largest impact on learning). Participants in this condition were given a larger base payment at the end of the study to match the average payment of the other conditions, in addition to any pre-/post-test performance bonuses.

Increases in difficulty level were not individually determined as they were in study 1 because the High-Certainty condition would, by definition, quickly reach a higher difficulty level at a faster pace than the other three training variations. Instead, all participants moved up in difficulty after a set number of blocks. For addition, participants completed 8 blocks of problems consisting of a single-digit number and a double-digit number (S/D blocks), and 22 blocks of problems consisting of two double-digit numbers (D/D blocks). For subtraction, participants completed 14 S/D blocks and 16 D/D blocks. The number of blocks was based on the median number of blocks that participants in Kallai et al.'s study (2011) needed to achieve 90 % accuracy at each difficulty level. The parameters of each difficulty level are shown in Table 1.

Complex math and number comparison tasks Both tasks were identical to those used in study 1, except that the complex math test was completed with paper-and-pencil and its completion time was collected.

Results

Training results

We verified the effectiveness of the four training variations by comparing accuracy on the first block of training with accuracy on the last block of training, as was done in study 1. First block and last block accuracies were compared separately for difficulty level 1 (D/S problems) and for difficulty level 2 (D/D problems), and separately for addition and subtraction problems. Accuracy was measured as exact solutions to the problems.

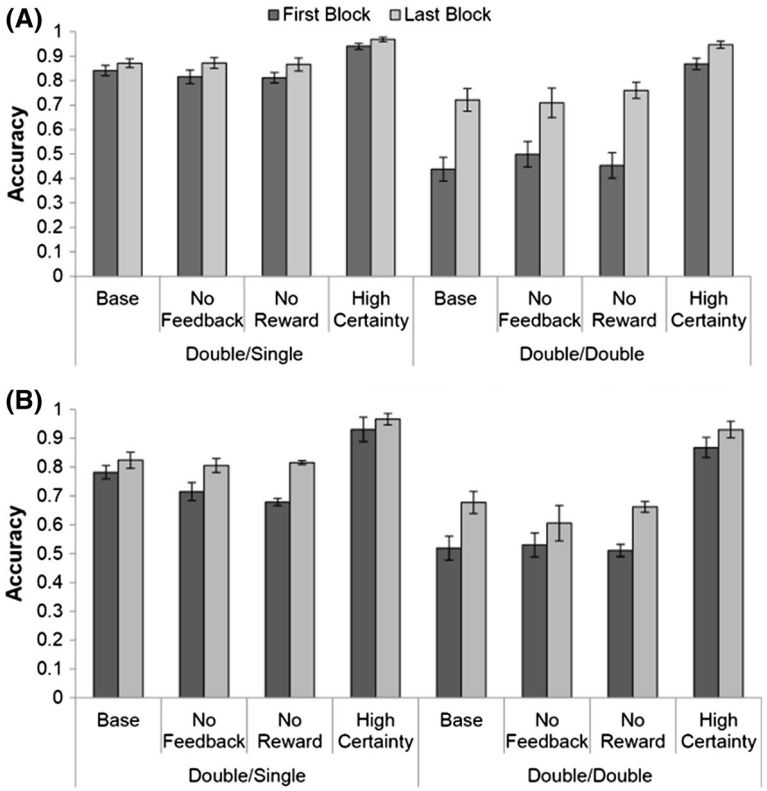

Fig. 5 Mean training accuracy scores (with SE bars) for the four training variations. Mean accuracy scores for (**a**) addition and (**b**) subtraction double/single and double/double training problems

Figure 5 shows the mean accuracy scores for the four training variations across the various problem types. As expected, the high certainty condition achieved higher accuracy than the other three training types from the beginning, and was consequently excluded from the condition effect analyses on the training data as there was not a comparable difficulty standard. However, despite starting at a higher accuracy than the other conditions, the High Certainty condition still improved significantly from pretest to post-test in all problem types [addition D/S: $t(19) = -3.27$, $p = 0.004$; addition D/D: $t(19) = -4.03$, $p = 0.001$; subtraction D/S: $t(19) = -2.80$, $p = 0.011$; subtraction D/D: $t(19) = -3.43$, $p = 0.003$].

Block (first, last) X Training Condition (Base, No Feedback, No Reward) repeated-measures ANOVAs found no significant accuracy differences between the other three conditions for any of the problem types [$F < 1$]. All problem types showed a significant main effect of Block, where last block accuracy was higher than first block accuracy [addition D/S: $F(1, 57) = 20.73$, $p < 0.001$, $\eta_p^2 = 0.27$; addition D/D: $F(1, 57) = 145.8$, $p < 0.001$, $\eta_p^2 = 0.72$; subtraction D/S: $F(1, 57) = 33.72$, $p < 0.001$, $\eta_p^2 = 0.37$; subtraction D/D: $F(1, 56) = 43.74$, $p < 0.001$, $\eta_p^2 = 0.44$]. Thus, all four training variations improved, and the three conditions that could be directly compared improved equally in arithmetic involving double-digit and single-digit problems.

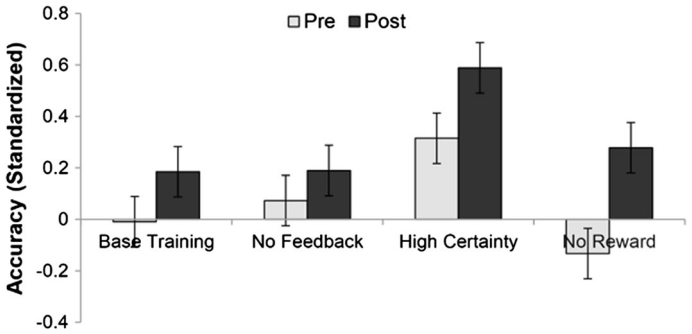

Fig. 6 Adjusted mean pre- to post-test gains in complex math standardized accuracy (with within SE bars) within each of the four training variations

Complex math

Complex math accuracy scores were standardized as they were in study 1. Both pre-test and post-test accuracy scores can be seen in Fig. 6. Although appearing to vary somewhat in the graph, there were no statistically significant differences between the four training conditions in pre-test accuracy [$F < 1$]. The Base, No-Feedback, High-Certainty, and No-Reward conditions correctly answered approximately 11, 11, 11, and 12 questions (as a mean out of 16), respectively, at pre-test.

Similarly to Study 1, we ran an ANCOVA on post-test Complex Math accuracy with the between-subject factor of Training Variation. We also included pre-test Complex Math accuracy and the gain in standardized completion times (standardized using the same process used to standardize accuracy scores) as covariates. There were no significant differences between Training Variations [$F(3, 74) = 0.359$, $p = 0.78$]. A Training Variation X Session (pre, post) repeated-measures ANOVA showed that only the main effect of Session was significant [$F(1, 75) = 6.89$, $p = 0.011$, $\eta_p^2 = 0.08$], such that post-test accuracy was higher than pre-test accuracy by an average of 0.25, replicating the prior transfer of learning results. In terms of questions answered, the Base, No-Feedback, High-Certainty, and No-Reward conditions correctly answered approximately 12, 12, 12, and 13 questions (as a mean out of 16), respectively (i.e., roughly showing a one question improvement on the test). The Training Variation and Session interaction was not significant [$F < 1$], suggesting that all training variations improved complex mathematical skills by a comparable amount.

Number comparison

To assess changes in numerical representation precision, we ran a Training Variation (Base, No-Feedback, High-Certainty, No-Reward) X Session (pre, post) X Distance (near, far) repeated-measures ANOVA on Number Comparison accuracy and response times, with Training Variation as a between-subjects factor, and Session and Distance as within-subject factors. Ratios were categorized using the same near/far distinction that was used in study 1. A Session X Distance interaction showed that accuracy significantly increased from pre-test to post-test for the near ratios (pre-test: $M = 0.96$, $SD = 0.02$; post-test: $M = 0.98$, $SD = 0.009$), but not the far ratios (pre-test: $M = 0.98$, $SD = 0.01$; post-test: $M = 0.98$, $SD = 0.009$) [$F(1, 76) = 26.1$, $p < 0.001$, $\eta_p^2 = 0.26$] (see Fig. 7), replicating

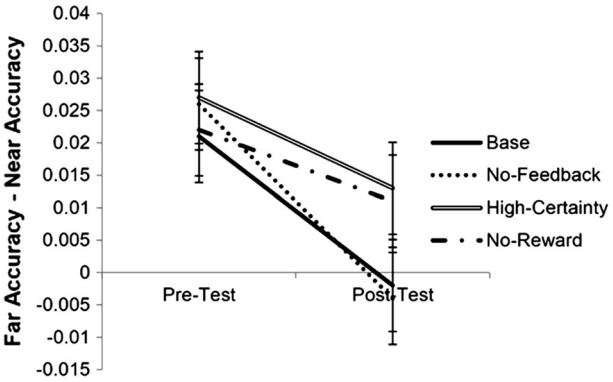

Fig. 7 Mean difference (and within SE bars) between far and near ratio accuracy on number comparison pre-test and post-test for the four training variations. Values closer to 0 indicate a smaller difference between far and near ratios

the core representational change after training from Study 1. The three-way interaction of Training Variation X Session X Distance was not significant, so this effect did not differ among the four training variations [$F(1, 76) = 1.36$, $p = 0.26$, $\eta_p^2 = 0.05$].

Meanwhile, for response time, participants were significantly faster for far ratios than near ratios [$F(1, 76) = 114.6$, $p < 0.001$, $\eta_p^2 = 0.60$]. There was no Distance X Group interaction [$F < 1$], but a marginal Distance X Session interaction was observed [$F(1, 76) = 3.75$, $p = 0.06$, $\eta_p^2 = 0.05$]. There was a significant Session X Group interaction [$F(1, 76) = 4.37$, $p = 0.007$, $\eta_p^2 = 0.15$], where the No-Feedback and High-Certainty variations responded more quickly at post-test than at pre-test and the No-Reward variation became slower at post-test.

Discussion

Study 2 compared three training variations (No Feedback, High Certainty, and No Reward) to the base training used in our previous studies. The results again showed the important transfer results in Complex Math and provided further evidence of underlying changes in representational precision. The gains in the complex math task and accuracy improvements in discriminating close symbolic quantities did not differ across the four versions of training. This suggests that the learning process of the mental computation training is robust, such that omitting one element from the design does not undermine its effectiveness.

The non-significant differences across the three training variations are somewhat inconsistent with literature on the impact of feedback, reward, and uncertainty on learning. For example, many studies show that corrective feedback confers great learning benefits (e.g., Anderson et al. 1971; Butler et al. 2008; Epstein et al. 2002; Metcalfe and Kornell 2007), but the lack of immediate feedback did not appear to affect the learning of the No-Feedback condition. However, it is debated whether immediate or delayed feedback leads to greater learning improvements (see Kulik and Kulik 1988). The No-Feedback condition may have still benefited from the delayed feedback given at the end of each training block.

Regarding the No-Reward condition, rewards are thought to benefit learning by positively reinforcing correct behaviors, but the lack of monetary rewards in the No-

Reward condition did not appear to have a strong effect. However, the motivational literature often draws a distinction between extrinsic motivation (i.e., motivation that comes from outside sources) and intrinsic motivation (i.e., motivation that comes from within an individual) (see Sansone and Harackiewicz 2000), with the latter related to higher academic success (Deci et al. 1991). While external rewards can potentially motivate additional engagement when they affirm competence at a task, extrinsic performance-contingent rewards (such as the monetary rewards used in training) have also been shown to undermine people's intrinsic motivation (see Deci et al. 1999 for a review). Removing extrinsic rewards may have led to a greater dependence on intrinsic motivation, leading to equal learning for the No-Reward condition.

Uncertainty is thought to benefit learning by increasing selective attention and engagement with uncertain stimuli (Fiorillo et al. 2003). People learn when the stimuli's predicted outcome mismatches its actual outcome, changing their behavior until their predictive quality improves and the uncertainty of outcomes decreases sufficiently (Dayan et al. 2000; Pearce and Hall 1980). Lack of uncertainty in the High-Certainty condition may have led to decreased attention toward the task and fewer opportunities for learning, but did not seem to have a large effect. Still, participants received more consistent positive feedback because of condition's lower difficulty, which may have acted as a task motivator. Furthermore, these participants received two forms of feedback—immediate trial-by-trial feedback and indirect reward feedback—that together may provide additional learning benefits.

Eliminating one core learning feature (immediate feedback, uncertainty, or reward) appears to have little effect on the training results. It is possible that eliminating two or more of these core learning features would have a greater detrimental effect on learning. The learning features may have overlapping benefits; for example, reward could be used as an indirect form of feedback (as participants are given a higher reward for more accurate performance), which could attenuate any effects caused by a lack of immediate feedback. Further experiments using training variations that remove two or more core learning features could determine whether a specific combination of features are required for the training to remain effective.

General discussion

Mathematical fluency is normally defined as the ability to perform mathematics quickly, accurately, and flexibly. In the past, fluency has been quantified as the number of math facts that can be quickly retrieved. While this measures the quick and accurate criteria of fluency, it neglects whether such math knowledge is also flexible; in several cases, knowledge gained through retrieval-based training has not transferred to non-trained problems (Bajic et al. 2011; Imbo and Vandierendonck 2008; Rickard et al. 1994). Alternative training programs have attempted to foster mathematical fluency through the development of meaningful representations of number, which involve a greater conceptual understanding of numbers and quantities and have the potential to be more flexible than retrieval-focused training programs.

In a study by Kallai et al. (2011), a training program that developed meaningful number representations through multi-digit mental computation led to marked improvements on a complex mathematics test that required skills beyond the addition and subtraction that was trained. The training also led to improvements on a symbolic representational precision task, with greater improvement on the complex math task correlating with greater improvement on the representational precision task. In the current report, we found that these

transfer and precision effects were robust. Significant improvements on the same complex math task were found when the training was completed by a diverse population outside of college students and recent graduates (study 1), and when isolated aspects of the original training were removed (study 2). There were no improvements for participants who did not undergo training. Similarly, in both studies, participants who completed the training showed signs of representational changes: study 1 training participants became faster and study 2 training participants became more accurate at distinguishing symbolic numerical quantities after training. Meanwhile, study 1 participants who did not take part in the training showed no improvement in their numerical precision, suggesting that the training process led to representational change. These results are similar to null results observed by Kallai et al. (2011) for a control training group and a simple test–retest group. Thus, it appears that mental computation training is responsible for both the transfer and representational change effects observed across studies, and that these effects are robustly obtained across populations and training variations.

The mental computation training likely affected more than just the representational acuity of double-digit numbers. It may also have improved addition and subtraction math facts, as well as procedural fluency in addition and subtraction. However, neither math facts nor procedural fluency can sufficiently explain the observed improvements on the complex math task. As suggested by previous research (Bajic et al. 2011; Imbo and Vandierendonck 2008; Rickard et al. 1994), math facts knowledge should only transfer to multi-digit arithmetic problems that are very similar to those used in the training task (e.g., $86 + 11 = ?$). No such problems were present in the complex math task. In addition, the training task used a large set of problems that were rarely repeated, which would make it difficult for participants to memorize and retrieve the facts for a later task. Procedural fluency could contribute to complex math gains, but its effects would be limited to problems that can be solved through addition and subtraction. The majority of the complex math task items involved more difficult procedures. Though some of the complex math problems could be restructured to rely primarily on addition or subtraction (e.g., plugging numbers into a number set problem), it would be considerably less efficient to use addition- or subtraction-dependent strategies compared to strategies based on conceptual understanding. In contrast, greater representational acuity would strengthen the connections between multi-digit numbers (used in the majority of the complex math problems) and their numerical quantities, providing a basis for greater conceptual and numerical understanding and making it a more likely mechanism for complex math improvements than math facts or procedural fluency alone.

Our results and proposed mechanism are consistent with previous literature that links mental computation ability to improved number sense (see Gersten and Chard 1999 for a review) and flexible number knowledge (e.g., Cobb and Merkel 1989; Klein and Beishuizen 1994; Maclellan 2001; Reys 1984; Sowder 1988). Thompson (1999) proposed that mental computation is developed through the contributions of different components, which include math facts (such as those fostered by retrieval-based training) and numerical understanding (e.g., properties of number relations), suggesting that mental computation proficiency requires strong number sense. Reciprocally, mental computation may also foster greater number sense because it makes properties of numbers and quantities more salient to students (Greeno 1991), and it may encourage students to consider the meaning of numbers and quantities within a problem, as opposed to following procedural steps without thinking about what they are doing (Maclellan 2001). While mental computation has been associated with improved number sense, its link to representational change has not been investigated as thoroughly. In the current studies, mental computation was

reinforced with learning features (immediate feedback, high uncertainty, and rewards) meant to encourage representational change, though we found that removing one of the learning features did not significantly impact representation precision improvements. We could gain a more detailed understanding of mental computation's effects on number representations by taking away more of these core learning features to see which (if any) are necessary for fostering representational change, or whether mental computation is enough.

Individuals who are skilled at mental computation also use a variety of mathematical strategies and are able to choose the most efficient strategy depending on the situation (e.g., Cooper et al. 1996; Heirdsfield and Cooper 2004; Sowder 1992). Several studies have increased the frequency of mental computation strategy use by directly teaching mental computation strategies or by encouraging students to generate their own strategies (see Varrol and Farran 2007 for a review), but they have been less successful in getting students to flexibly use different mathematical strategies. For example, Klein and Beishuizen (1994) observed classrooms in which a specific mental computation strategy was taught. Even though flexibility was encouraged, the majority of students preferred to use the taught strategy, implying that students may not attempt to flexibly use other strategies if they are taught a specific method. Even when students generate their own strategies, students often limit themselves to a mental version of an already-learned algorithm (Reys et al. 1995), suggesting that early emphasis on written algorithms may discourage students from developing their own mental computation strategies (Hope 1987). In our mental computation training, mental computation is encouraged through the training design, but no specific computation strategy is given to participants. In future studies, it would be interesting to investigate whether our training supports the generation of flexible mental computation strategies or strategies based on known algorithms, and whether individual differences in strategies influence the transfer of mathematical knowledge.

Although mathematical fluency is defined as mathematical performance that is quick, accurate, and flexible, retrieval-based training programs have primarily used speed and accuracy as a measure of fluency. The current studies suggest that training programs that focus on the development of numerical representations can not only improve the speed and accuracy of arithmetic computation, they can also robustly lead to flexible and transferrable gains in complex mathematical performance. In conclusion, training programs built on numerical representations can lead to a mathematical fluency that is consistently quick, accurate, and flexible, providing an alternative and effective basis for fostering advanced mathematical knowledge.

Acknowledgments This research was supported by NSF 0815945 from the National Science Foundation and by Award Number T32GM081760 from the National Institute of General Medical Sciences. The authors would also like to thank Michael Skirpan for his help in programming the tasks used on Amazon Mechanical Turk.

References

Anderson, R. C., Kulhavy, R. W., & Andre, T. (1971). Feedback procedures in programmed instruction. *Journal of Educational Psychology, 62*(2), 148–156. doi:10.1037/h0030766.

Aron, A. R., Shohamy, D., Clark, J., Myers, C., Gluck, M. A., & Poldrack, R. A. (2004). Human midbrain sensitivity to cognitive feedback and uncertainty during classification learning. *Journal of Neurophysiology, 92*, 1144–1152. doi:10.1152/jn.01209.2003.

Bajic, D., Kwak, J., & Rickard, T. C. (2011). Specificity of learning through memory retrieval practice: The case of addition and subtraction. *Psychonomic Bulletin & Review, 18*(6), 1148–1155. doi:10.3758/s13423-011-0151-4.

Berns, G. S., McClure, S. M., Pagnoni, G., & Montague, P. R. (2001). Predictability modulates human brain response to reward. *The Journal of Neuroscience, 21*(8), 2793–2798.

Buhrmester, M., Kwang, T., & Gosling, S. D. (2011). Amazon's Mechanical Turk a new source of inexpensive, yet high-quality, data? *Perspectives on Psychological Science, 6*(1), 3–5.

Butler, A. C., Karpicke, J. D., & Roediger, H. L. (2008). Correcting a metacognitive error: Feedback increases retention of low-confidence correct responses. *Journal of Experimental Psychology. Learning, Memory, and Cognition, 34*(4), 918–928. doi:10.1037/0278-7393.34.4.918.

Cobb, P., & Merkel, G. (1989). Thinking strategies: Teaching arithmetic through problem solving. In P. R. Trafton (Ed.), *New directions for elementary school mathematics* (pp. 70–81). Reston: National Council of Teachers of Mathematics.

Cooper, T. J., Heirdsfield, A., & Irons, C. J. (1996). Children's mental strategies for addition and subtraction word problems. In J. Mulligan & M. Mitchelmore (Eds.), *Children's number learning* (pp. 147–162). Adelaide: Australian Association of Mathematics Teachers Inc.

Dayan, P., Kakade, S., & Montague, P. R. (2000). Learning and selective attention. *Nature, 3*, 1218–1223. doi:10.1038/81504.

Deci, E. L., Koestner, R., & Ryan, R. M. (1999). A meta-analytic review of experiments examining the effects of extrinsic rewards on intrinsic motivation. *Psychological Bulletin, 125*(6), 627–668.

Deci, E. L., Vallerand, R. J., Pelletier, L. G., & Ryan, R. M. (1991). Motivation and education: The self-determination perspective. *Educational Psychologist, 26*(3), 325–346. doi:10.1207/s15326985ep2603&4_6.

Dehaene, S. (1992). Varieties of numerical abilities. *Cognition, 44*, 1–42. doi:10.1016/0010-0277(92)90049-N.

Dehaene, S., Bossini, S., & Giraux, P. (1993). The mental representation of parity and number magnitude. *Journal of Experimental Psychology: General, 122*(3), 371–396. doi:10.1037/0096-3445.122.3.371.

Dehaene, S., Dupoux, E., & Mehler, J. (1990). Is numerical comparison digital? Analogical and symbolic effects in two-digit number comparison. *Journal of Experimental Psychology: Human Perception and Performance, 16*(3), 626–641. doi:10.1037/0096-1523.16.3.626.

Dehaene, S., & Marques, J. F. (2002). Cognitive euroscience: Scalar variability in price estimation and the cognitive consequences of switching to the euro. *Quarterly Journal of Experimental Psychology, 55*(3), 705–731. doi:10.1080/02724980244000044.

Dehaene, S., Piazza, M., Pinel, P., & Cohen, L. (2003). Three parietal circuits for number processing. *Cognitive Neuropsychology, 20*(3), 487–506. doi:10.1080/02643290244000239.

Delaney, H. D., & Maxwell, S. E. (1981). On using analysis of covariance in repeated measures designs. *Multivariate Behavioral Research, 16*(1), 105–123.

Delazer, M., Domahs, F., Bartha, L., Brenneis, C., Lochy, A., Trieb, T., & Benke, T. (2003). Learning complex arithmetic—an fMRI study. *Cognitive Brain Research, 18*(1), 76–88. doi:10.1016/j.cogbrainres.2003.09.005.

Delgado, M. R., Stenger, V. A., & Fiez, J. A. (2004). Motivation-dependent responses in the human caudate nucleus. *Cerebral Cortex, 14*, 1022–1030. doi:10.1093/cercor/bhh062.

Dimitrov, D. M., & Rumrill, P. D, Jr. (2003). Pretest-posttest designs and measurement of change. *Work: A Journal of Prevention, Assessment and Rehabilitation, 20*(2), 159–165.

Epstein, M. L., Lazarus, A. D., Calvano, T. B., Matthews, K. A., Hendel, R. A., Epstein, B. B., & Brosvic, G. M. (2002). Immediate feedback assessment technique promotes learning and corrects inaccurate first responses. *The Psychological Record, 52*, 187–201.

Fiorillo, C. D., Tobler, P. N., & Schultz, W. (2003). Discrete coding of reward probability and uncertainty by dopamine neurons. *Science, 299*(5614), 1898–1902. doi:10.1126/science.1077349.

Fischer, U., Moeller, K., Bientzle, M., Cress, U., & Nuerk, H.-C. (2011). Sensori-motor spatial training of number magnitude representation. *Psychonomic Bulletin & Review, 18*(1), 177–183. doi:10.3758/s13423-010-0031-3.

Gallistel, C. R., & Gelman, R. (1992). Preverbal and verbal counting and computation. *Cognition, 44*, 43–74. doi:10.1016/0010-0277(92)90050-R.

Gallistel, C. R., & Gelman, R. (2005). Mathematical cognition. In K. J. Holyoak & R. G. Morrison (Eds.), *The Cambridge handbook of thinking and reasoning* (pp. 559–588). New York: Cambridge University Press. doi:10.1037/002775.

Gersten, R., & Chard, D. J. (1999). Number sense: Rethinking arithmetic instruction for students with mathematical disabilities. *The Journal of Special Education, 33*, 18–28. doi:10.1177/002246699903300102.

Gilmore, C. K., McCarthy, S. E., & Spelke, E. S. (2010). Non-symbolic arithmetic abilities and achievement in the first year of formal schooling in mathematics. *Cognition, 115*(3), 394–406. doi:10.1016/j.cognition.2010.02.002.

Goodman, J. K., Cryder, C. E., & Cheema, A. (2013). Data collection in a flat world: The strengths and weaknesses of Mechanical Turk samples. *Journal of Behavioral Decision Making, 26*(3), 213–224.

Greeno, J. G. (1991). Number sense as situated knowing in a conceptual domain. *Journal for Research in Mathematics Education, 22*(3), 170–218.

Halberda, J., Ly, R., Wilmer, J. B., Naiman, D. Q., & Germine, L. (2012). Number sense across the lifespan as revealed by a massive Internet-based sample. *Proceedings of the National Academy of Sciences, 109*(28), 11116–11120. doi:10.1073/pnas.1200196109.

Halberda, J., Mazzocco, M. M. M., & Feigenson, L. (2008). Individual differences in non-verbal number acuity correlate with maths achievement. *Nature, 455*, 665–668. doi:10.1038/nature07246.

Heirdsfield, A. M., & Cooper, T. J. (2004). Factors affecting the process of proficient mental addition and subtraction: Case studies of flexible and inflexible computers. *The Journal of Mathematical Behavior, 23*(4), 443–463.

Holloway, I. D., & Ansari, D. (2009). Mapping numerical magnitudes onto symbols: The numerical distance effect and individual differences in children's mathematics achievement. *Journal of Experimental Child Psychology, 103*(1), 17–29. doi:10.1016/j.jecp.2008.04.001.

Hope, J. A. (1987). A case study of a highly skilled mental calculator. *Journal for Research in Mathematics Education, 18*(5), 331–342.

Imbo, I., Duverne, S., & Lemaire, P. (2007). Working memory, strategy execution, and strategy selection in mental arithmetic. *Quarterly Journal of Experimental Psychology, 60*(9), 1246–1264. doi:10.1080/17470210600943419.

Imbo, I., & Vandierendonck, A. (2008). Practice effects on strategy selection and strategy efficiency in simple mental arithmetic. *Psychological Research, 72*(5), 528–541. doi:10.1007/s00426-007-0128-0.

Inglis, M., & Gilmore, C. (2014). Indexing the approximate number system. *Acta Psychologica, 145*, 147–155. doi:10.1016/j.actpsy.2013.11.009.

Ischebeck, A., Zamarian, L., Egger, K., Schocke, M., & Delazer, M. (2007). Imaging early practice effects in arithmetic. *NeuroImage, 36*, 993–1003. doi:10.1016/j.neuroimage.2007.03.051.

Ischebeck, A., Zamarian, L., Schocke, M., & Delazer, M. (2009). Flexible transfer of knowledge in mental arithmetic–an fMRI study. *NeuroImage, 44*(3), 1103–1112. doi:10.1016/j.neuroimage.2008.10.025.

Kallai, A. Y., Schunn, C. D., Ponting, A. L., & Fiez, J. A. (2011). Improving foundational number representations through simple arithmetical training. *Society for Research on Educational Effectiveness*. Evanston: Society for Research on Educational Effectiveness.

Klein, T., & Beishuizen, M. (1994). Assessment of flexibility in mental arithmetic. In J. E. H. LuitVan (Ed.), *Research and instruction in kindergarten and primary school* (pp. 125–152). Doetinchem: Graviatt Publishing Company.

Kulik, J. A., & Kulik, C.-L. C. (1988). Timing of feedback and verbal learning. *Review of Educational Research, 58*(1), 79–97. doi:10.3102/00346543058001079.

Lourenco, S. F., Bonny, J. W., Fernandez, E. P., & Rao, S. (2012). Nonsymbolic number and cumulative area representations contribute shared and unique variance to symbolic math competence. *Proceedings of the National Academy of Sciences, 109*(46), 18737–18742. doi:10.1073/pnas.1207212109.

Maclellan, E. (2001). Mental calculation: Its place in the development of numeracy. *Westminster Studies in Education, 24*(2), 145–154. doi:10.1080/0140672010240205.

Markovits, Z., & Sowder, J. (1994). Developing number sense: An intervention study in grade 7. *Journal for Research in Mathematics Education, 25*(1), 4–29. doi:10.2307/749290.

Mason, W., & Suri, S. (2012). Conducting behavioral research on Amazon's Mechanical Turk. *Behavior Research Methods, 44*(1), 1–23.

Mazzocco, M. M. M., Devlin, K. T., & McKenney, S. J. (2008). Is it a fact? Timed arithmetic performance of children with mathematical learning disabilities (MLD) varies as a function of how MLD is defined. *Developmental Neuropsychology, 33*(3), 318–344. doi:10.1080/87565640801982403.

Metcalfe, J., & Kornell, N. (2007). Principles of cognitive science in education: The effects of generation, errors, and feedback. *Psychonomic Bulletin & Review, 14*(2), 225–229. doi:10.3758/BF03194056.

Moyer, R. S., & Landauer, T. K. (1967). Time required for judgements of numerical inequality. *Nature, 215*, 1519–1520. doi:10.1038/2151519a0.

National Council of Teachers of Mathematics. (2000). *Principles and Standards for School Mathematics* (3rd ed., p. 20). Reston: National Council of Teachers of Mathematics.

National Governors Association Center for Best Practices & Council of Chief State School Officers. (2010). In National Governors Association Center for Best Practices & Council of Chief State School Officers

(Ed.) *Common core state standards for mathematics*. Washington, DC. Retrieved from https://www.corestandards.org/wp-content/uploads/Math_Standards.pdf.
Obersteiner, A., Reiss, K., & Ufer, S. (2013). How training on exact or approximate mental representations of number can enhance first-grade students' basic number processing and arithmetic skills. *Learning and Instruction, 23*, 125–135. doi:10.1016/j.learninstruc.2012.08.004.
Paolacci, G., Chandler, J., & Ipeirotis, P. G. (2010). Running experiments on Amazon Mechanical Turk. *Judgment and Decision Making, 5*(5), 411–419.
Park, J., & Brannon, E. M. (2013). Training the approximate number system improves math proficiency. *Psychological Science, 24*(10), 2013–2019. doi:10.1177/0956797613482944.
Pearce, J. M., & Hall, G. (1980). A model for Pavlovian learning: Variations in the effectiveness of conditioned but not of unconditioned stimuli. *Psychological Review, 87*(6), 532–552.
Peer, E., Vosgerau, J., & Acquisti, A. (2014). Reputation as a sufficient condition for data quality on Amazon Mechanical Turk. *Behavioral Research, 46*, 1023–1031. doi:10.3758/s13428-013-0434-y.
Pellegrino, J. W., & Goldman, S. R. (1987). Information processing and elementary mathematics. *Journal of Learning Disabilities, 20*(1), 23–32. doi:10.1177/002221948702000105.
Price, G. R., Mazzocco, M. M. M., & Ansari, D. (2013). Why mental arithmetic counts: Brain activation during single digit arithmetic predicts high school math scores. *The Journal of Neuroscience, 33*(1), 156–163. doi:10.1523/JNEUROSCI.2936-12.2013.
Ramani, G. B., & Siegler, R. S. (2011). Reducing the gap in numerical knowledge between low- and middle-income preschoolers. *Journal of Applied Developmental Psychology, 32*, 146–159. doi:10.1016/j.appdev.2011.02.005.
Resnick, L. B. (1983). Mathematics and science learning: A new conception. *Science, 220*(4596), 477–478. doi:10.1126/science.220.4596.477.
Reys, R. E. (1984). Mental computation and estimation: Past, present, and future. *The Elementary School Journal, 84*(5), 546–557.
Reys, R. E., Reys, B. J., Nohda, N., & Emori, H. (1995). Mental computation performance and strategy use of Japanese students in grades 2, 4, 6, and 8. *Journal for Research in Mathematics Education, 26*(4), 304–326.
Rickard, T. C., Healy, A. F., & Bourne, L. E, Jr. (1994). On the cognitive structure of basic arithmetic skills: Operation, order, and symbol transfer effects. *Journal of Experimental Psychology. Learning, Memory, and Cognition, 20*(5), 1139–1153. doi:10.1037/0278-7393.20.5.1139.
Salthouse, T. A. (2010). Influence of age on practice effects in longitudinal neurocognitive change. *Neuropsychology, 24*(5), 563–572. doi:10.1037/a0019026.
Sansone, C., & Harackiewicz, J. M. (2000). Intrinsic and extrinsic motivation: The search for optimal motivation and performance. In C. Sansone & J. M. Harackiewicz (Eds.), *Intrinsic and extrinsic motivation: The search for optimal motivation and performance* (p. 489). San Diego: Academic Press.
Scholastic Inc. (2005). Research foundation & evidence of effectiveness for FASTT Math. Retrieved from http://research.scholastic.com/.
Schunn, C. D., Reder, L. M., Nhouyvanisvong, A., Richards, D. R., & Stroffolino, P. J. (1997). To calculate or not to calculate: A source activation confusion model of problem familiarity's role in strategy selection. *Journal of Experimental Psychology. Learning, Memory, and Cognition, 23*(1), 3–29.
Simcox, T., & Fiez, J. A. (2013). Collecting response times using Amazon Mechanical Turk and Adobe Flash. *Behavior Research Methods,*. doi:10.3758/s13428-013-0345-y.
Sowder, J. (1988). Mental computation and number comparisons: Their roles in the development of number sense and computational estimation. In J. Hiebert & M. Behr (Eds.), *Number concepts and operations in the middle grades* (pp. 182–197). Hillsdale: National Council of Teachers of Mathematics.
Sowder, J. T. (1992). Making sense of numbers in school mathematics. In G. Leinhardt, R. T. Putnam, & R. A. Hattrup (Eds.), *Analysis of arithmetic for mathematics teaching* (pp. 1–51). Hillsdale: Lawrence Erlbaum Associates Inc.
Sweller, J., Mawer, R. F., & Ward, M. R. (1983). Development of expertise in mathematical problem solving. *Journal of Experimental Psychology: General, 112*(4), 639–661. doi:10.1037/0096-3445.112.4.639.
Thompson, I. (1999). Getting your head around mental computation. In I. Thompson (Ed.), *Issues in teaching numeracy in primary schools* (pp. 145–156). Buckingham: Open University Press.
Tricomi, E. M., Delgado, M. R., & Fiez, J. A. (2004). Modulation of caudate activity by action contingency. *Neuron, 41*, 281–292.
Tricomi, E., & Fiez, J. A. (2008). Feedback signals in the caudate reflect goal achievement on a declarative memory task. *NeuroImage, 41*, 1154–1167. doi:10.1016/j.neuroimage.2008.02.066.
Van Breukelen, G. J. (2006). ANCOVA versus change from baseline had more power in randomized studies and more bias in nonrandomized studies. *Journal of Clinical Epidemiology, 59*(9), 920–925.

Varrol, F., & Farran, D. (2007). Elementary school students' mental computation proficiencies. *Early Childhood Education Journal, 35*(1), 89–94. doi:10.1007/s10643-007-0173-8.

Whyte, J. C., & Bull, R. (2008). Number games, magnitude representation, and basic number skills in preschoolers. *Developmental Psychology, 44*(2), 588–596. doi:10.1037/0012-1649.44.2.588.

Woodward, J. (2006). Developing automaticity in multiplication facts: Integrating strategy instruction with timed practice drills. *Learning Disability Quarterly, 29*, 269–289. doi:10.2307/30035554.