

Music Training and Mathematics Achievement:  
A Multi-Year, Iterative Project Designed to Enhance Student Learning

Paper presented at:  
the Annual Conference of the American Psychological Association  
Washington, DC, August, 2005

To appear in:  
Kelly, A.E. & Lesh, R. A. (Eds.), Handbook of design research in mathematics, science,  
and technology education.

Michael E. Martinez<sup>3</sup>, Matthew Peterson<sup>1,2</sup>, Mark Bodner<sup>1,3,4</sup> Andrew  
Coulson<sup>1</sup>, Sydni Vuong<sup>1</sup>, Wenjie Hu<sup>1</sup>, Tina Earl<sup>1</sup>, Jill S. Hansen<sup>1,3</sup>, & Gordon L.  
Shaw<sup>1,3</sup>

1. MIND Institute, Costa Mesa, CA
2. University of California, Berkeley
3. University of California, Irvine
4. University of California, Los Angeles

## Music Training and Mathematics Achievement:

### A Multi-Year, Iterative Project Designed to Enhance Student Learning

In a seminal study of the Mozart effect, college students who listened to the first ten minutes of the Mozart Sonata for Two Pianos in D Major (K.448) experienced a short-term enhancement of spatial-temporal (ST) reasoning (Rauscher, Shaw, & Ky, 1993; 1995). Findings from this study were widely reported in the popular media. The community of research psychologists took a more critical view of the study and its implications than did the general public. Indeed, among psychologists and educational researchers, the Mozart effect has been highly controversial. A claimed connection between listening to Mozart and becoming smarter as a result seemed scarcely believable, more akin to magic than to a cognitive effect. Moreover, the results of the original study, when replicated, sometimes seemed explainable by factors not originally accounted for, including arousal or musical preference.

Over time, the original claims of the Mozart effect have broadened to de-emphasize short-term improvement of cognitive functioning. At the same time, empirical studies and theoretical developments have lent credibility and coherence to claims about associations between musical experience and certain forms of cognition. This nexus takes the form of neuronal circuitry common to (1) music listening and production, (2) mathematical cognition, and (3) spatial-temporal reasoning. In relation to this, Schellenberg (2004) showed significant increases in measured IQ with music instruction. These associations, sometimes

called the Generalized Mozart Effect (Shaw, 2000), are detectable on multiple measures: behavioral and neurophysiological data, fMRI imaging (Bodner et al., 2001), treatment of epilepsy patients (Hughes et al., 1998, 1999), EEG (Sarthein et al., 1997), animal models (Rauscher et al., 1998), music training studies (Rauscher et al., 1997; Graziano et al., 1999; Schellenberg, 2004), computational/theoretical models (Leng & Shaw, 1991; Bodner & Shaw, 2001), and meta-analyses (Hetland, 2000).

During the past six years, a research and development team at the MIND Institute in Costa Mesa, California, has tested the possibility that a combination of musical and spatial-temporal (ST) training can promote mathematics learning among elementary school students. The pedagogical approach was conceptualized to be distinct from standard ways of teaching mathematics. Typically, mathematics instruction relies heavily on symbolic notation in the form of numerals, operations, and equations. Often, and more fundamentally, these formalisms express patterns that can be represented as images or transformations of images. Whereas the formalisms of symbolic mathematics are ultimately conventions, the underlying patterns of mathematics are expressions of the natural or experienced world. The ability to make sense of patterns is, in our research program, assumed to be an inherent human capacity. The human brain has an innate ability to find and manipulate patterns. This pattern-finding capacity, largely experienced as subjective imagery, is a natural, near-universal propensity of the human mind and its underlying neural circuitry. Self-reports of many eminent scientists clarify that

their most creative notions often arise from ideas experienced as visual mental images. Hawking, Feynmann, Einstein and countless other luminaries have reported that their intellectual breakthroughs were at least sometimes experienced as dynamic mental images. Only later are these intuitions reduced to the precise language of equations.

In the research reported here, certain experiences with music are believed to evoke activation of brain circuitry common also to spatial-temporal reasoning. Huttenlocher (2002) noted that “spatial tasks and music are both represented in the same general cortical region, the non-dominant parietal cortex posterior to the postcentral gyrus” (p. 161). Huttenlocher raised the possibility that brain activity associated with one task (e.g., music interpretation) might “prime” the performance of another cognitive task whose associated brain activity is anatomically proximal (e.g., spatial-temporal reasoning). Data from brain imaging documents reveal correspondences between the two. For more than a decade now, numerous brain imaging studies have evidenced robust anatomical correspondences between brain areas activated by certain musical forms and by tasks that require spatial-temporal reasoning. Listening to the Mozart sonata (k. 448), for example, has been demonstrated in imaging studies to activate a brain network that includes:

- The dorsolateral prefrontal cortex (Brodmann areas 9 and 46)
- Specific areas of occipital lobe (Brodmann areas 17, 18 & 19)  
whose activation has been implicated in visual representation

- The cerebellum, implicated in the manipulation of visual representations (e.g., rotation)

These same areas become activated when subjects engage in spatial-temporal tasks, such as the Stanford-Binet paper folding task (Muftuler et al., 2004). Specific areas activated such as dorsolateral prefrontal cortex are known to be important in coordinating complex tasks in working memory (Bodner et al., 1996; Fuster, 2000). These areas were not always activated when listeners experienced other musical compositions, such as Beethoven's *Fur Elise*. Inter-subject variability greatly complicates the task of identifying specific cognitive functions to particular brain circuits. The brain activation patterns described here were notable for their cross-subject consistency.

The fact of common neural circuitry supporting spatial-temporal cognition and the experience of particular musical forms led to the conjecture that music could enhance spatial-temporal cognition. There was empirical support for this priming function. In one study, preschool children who received piano keyboard training for six months made substantial gains on a spatial-temporal reasoning task (Rauscher et al., 1997). Contrasting control groups did not improve significantly on the ST task. If music could enhance ST cognition, a second conjecture was that the same music might also enhance mathematics learning—if mathematics were taught and learned in a spatial-temporal mode. Specifically, the experience of some musical forms could prime brain circuitry that would support learning mathematical patterns depicted as transformation of images.

The following chapter traces the MIND Institute's project, called M+M (Math+Music) during the academic years 1998 through 2004. Our intent is, first, to describe the intervention and its effects. A second goal is to show how the M+M project evolved through feedback over the course of its implementation. Recursive cycles of development, intervention, and redesign of the intervention illustrate one example of the research genre called design experiments. Design experiments have been defined as “extended (iterative), interventionist (innovative and design-based), and theory-oriented enterprises whose “theories” do real work in practical educational contexts (Cobb, Confrey, diSessa, Lehrer, & Schauble, 2003; Cobb & Gravemeijer, this volume). Real-life educational contexts are, in turn, settings for experimental tests of the intervention. In path-breaking work on the design experiment methodology, Brown (1992) attempted to “engineer innovative educational environments and simultaneously conduct studies of those innovations” (p. 141). The M+M project satisfies these criteria in that the intervention proceeds from and informs a general theoretical model. Like all design experiments, it is highly contextualized in real-world settings (mostly, classrooms), is implemented over several years, and is iterative in using feedback from one intervention, “draft n,” to inform the design of draft “n + 1”. During the course of this project, many project aspects evolved simultaneously. These included the intervention proper, but also the nature and range of feedback, the length of the feedback cycle, and the underlying local and general theories motivating and guiding the project.

Our general conclusion is that elementary school students trained in piano keyboard and spatial-temporal reasoning significantly outperformed control group students on standardized measures of mathematical achievement. These data support hypotheses about cognitive and neurological associations between music and spatial-temporal reasoning, and affirm their combined ability to enhance learning in mathematics. Our own experiences also illuminate how the efforts of a group of researchers, over several years of sustained effort, can produce revisions and refinements to a guiding theory, and can result in progressively more effective experiences for learners.

## Method

### Students

M+M has been tested on students from grades K through 4, but in this chapter we focus mainly on second graders. M+M was developed originally for second graders, and the project staff has the most complete data from these students. Most participating second graders were primarily from low-SES backgrounds. We further concentrate on data from schools in which some second graders did not participate in the M+M program. Both participating and control classes exhibited a range of academic achievement and did not differ systematically in their demographic composition.

### Instructional Components

Keyboard Training. The music component of M+M was designed to teach students some basic musical concepts and skills necessary for playing the piano keyboard. One reason for selecting the piano keyboard for teaching musical

skills is that a keyboard allows the user to see the entire instrument while performing—all keys are in plain view and spatially correlated to the musical staff. The keyboard also avoids the need for excessively technical manual dexterity as is required in many string instruments. It is possible for a beginning keyboard player to sound clear notes and simple melodies almost from the start of instruction.

Normally, students received two lessons per week in keyboard instruction. A typical 45-minute lesson consisted of an initial segment of such activities as clapping note values, reading musical notation, learning about composers, and active listening to the first movement of the Mozart Sonata (K. 448). During keyboard practice, children sometimes worked independently and at other times in groups to learn a repertoire of progressively more difficult pieces. By the end of second grade, participating children had developed skill in a variety of musical activities. These included the manipulation of note values as small as sixteenths, knowledge of basic musical symbols, and recognition of the basic sonata form in listening examples. All students mastered approximately 15 pieces that employed parallel and contrary motion between the hands, as well as textures involving right-hand melody with left-hand accompaniment.

STAR software. A second component of M+M is training in mathematical concepts via spatial-temporal reasoning. This training was accomplished through a series of computer-based experiences using software named STAR. STAR (spatial-temporal animated reasoning) software was designed to develop skill in transforming mental images and so, presumably, to enhance spatial-

temporal abilities generally. The transformations involve symmetry operations applied to two-dimensional figures: folding and unfolding around multiple axes in the x-y plane of the computer screen, rotations around the z axis perpendicular to the screen, 180 degree flips around the x and y axes, and translations in the x-y plane.

Other activities challenge children to apply their spatial-temporal skills to solve mathematics problems, in particular, problems involving fractions, proportions, and symmetries. These are the kinds of mathematical operations that are difficult to teach using the language-analytic approach, which relies heavily on words and symbols. They are more amenable to visual image representations.

The STAR software presents a series of games for each mathematics topic. Each game computes a current score to indicate student progress. Students are required to display mastery on a level before progressing to a more difficult level within a game. The software tracked all actions taken by the learner as well as the scores attained.

### Mathematics Achievement

We report Stanford 9 and CAT 6 mathematics scores as indicators of general mathematics achievement. We also report program effects by results on the California Standards Test, a criterion-reference test that indexes students' success in mastering prescribed content standards. These tests were used as standardized instruments of accountability in the California public schools. In addition to these broad measures of mathematics achievement, the

project staff sometimes also used their own tests designed to measure specific aspects of mathematics learning.

## Results

In this section, findings are reported year by year, starting with pilot work in 1997. We note the findings for each year and the effect of those findings on subsequent redesign of our intervention.

### Preliminary Work: 1997-1998

Starting in 1997, preliminary work involved two studies, a pilot study and a main study (Graziano, Peterson, & Shaw, 1999). In the pilot study, second grade students engaged in computer-based exercises designed to enhance ST skill. The experimental design involved three groups: Group 1 (n=19) engaged the ST computer game exercises, Group 2 (n=20) used a computer game that taught English language skills; Group 3 (n=62) received no extra instruction. Group comparisons showed that Group 1 students scored significantly higher in ST mathematics reasoning in contrast with both Group 2 and Group 3.

The main study, also involving second graders, tested the effects of piano keyboard training paired with the ST computer games on spatial-temporal reasoning. The main experimental group (Piano-ST, n=26) was contrasted with a second group (English-ST, n=29), who received computer-based instruction in English language skills. A third group (No Lesson, n=28) received no additional instruction. On a computer-administered post-test of such math concepts as fractions and proportionality, the Piano-ST group showed significantly higher outcomes ( $p < .05$ ) than the other two comparison groups.

### 1998-1999

In the first year of project implementation, the M+M intervention was tested in a single urban elementary school. The centerpiece of the intervention was a version of the computer game software used in pilot work to enhance ST reasoning. Students who engaged in the computer exercises improved their performance on the Stanford 9 mathematics subtest. On the Stanford 9, M+M second graders' (n=18) average performance was at the 65th national percentile. A comparison group of second graders (n=36) averaged at the 36th percentile.

### 1999-2000

Starting in the fall of 1999, the project expanded from one elementary school to four. One lesson learned from earlier versions of M+M is that a desirable design feature of the software—concentrating on visual image-based representations—was probably taken too far. To be clear, the studied attempt to completely avoid mathematical notation and language was a miscalculation. Project staff realized that children needed to use both verbal and mathematical/symbolic language to express their mathematical ideas. In response, the project introduced a new teacher-led component. This new element, called “math integration,” bridged students’ experience of the – computer game software (and their resulting ST learning) to standard language-analytic expressions of mathematics, including numerals and symbolic notation for operations.

To some degree, the addition of a “math integration” component was a concession to the known requirements of tests—that assessment of mathematic achievement inevitably requires a student to use conventional symbols and language. Although students demonstrated mathematical competence in the symmetry and proportionality operations required by the software games, they were less competent to translate their ST representations into the symbolic notation and forms required on standardized tests.

Other insights were gained during this year. Not all of these insights can be tied to strictly controlled research designs and to reliable measurement of valid learning constructs. For example, the research team became aware that transfer of training was a problem. As psychologists and educators have learned and relearned through the decades, beginning with E. L. Thorndike, it is very easy to overestimate the applicability of learned skills to new problem contexts. Many skills that students ostensibly performed competently did not transfer to other tasks that required essentially the same skill set. Rather, students needed much more direct assistance, or scaffolding, to accomplish this transfer. The response was to train students more directly to transfer their newly-acquired concepts and skills to new tasks. Particularly beguiling was students’ use of mathematical language. The use of the right words suggested that students understood the relevant concept, but this was often not the case. Often, a lack of understanding was revealed by students’ difficulty in crossing between language-analytic representations and spatial-temporal

problems. It was not uncommon for a student to show competence in one modality, but be unable to translate that competence to another expressive form.

Another rather basic lesson that had to be learned and relearned was that *showing* students was not enough to ensure understanding. It was necessary to ask them to *do* something. A compelling visual illustration showing that a triangle's area was exactly half of its extension to a parallelogram did not, of course, translate directly to the formula for computing the area of a triangle,  $1/2 \text{ base} \times \text{height}$ . Such are the simple lessons, or rather fundamental assumptions, of constructivism. Yet design experiment research can, through hard experience, bring us back to foundational ideas.

### 2000-2001

In the third year of project implementation, three elementary schools had both participating and nonparticipating students. The summative comparisons of participating and control students showed an advantage for the M+M experience. Students in the M+M intervention had a higher mean score on the Stanford 9 Mathematics test ( $M=55$ ,  $SD=11.2$ ,  $N=102$ ) than did nonparticipating control group students ( $M=40.2$ ,  $SD=17.8$ ,  $N=77$ ). In each of the schools, M+M students outperformed control group students on the Stanford 9 Mathematics test ( $p<0.2$ ), but not to a degree that was statistically significant by the most common criteria. In this year also, the project staff appreciated more deeply the connections between language and measures of

assessment. The language integration portion of the M+M program continued to be important.

M+M -made regular attempts to send information to teachers about the progress of individual students. The form of this feedback shifted over time in accord with changing technologies, project team expertise, and experience with what was effective. During this project year, class- and student-level feedback was sent to teachers by conventional mail. Many teachers did not have e-mail at this time, so the low-tech option was chosen even with its obvious disadvantages of time lag.

### 2001-2002

During this academic year, M+M spread to seven schools that had both participating and nonparticipating students. A contrast of treatment and control groups on the Stanford 9 Mathematics Test showed a significant advantage of M+M students ( $M=63.7$ ,  $SD=21$ ,  $N=514$ ) over control group students ( $M=48.8$ ,  $SD=21$ ,  $N=285$ ). This difference was statistically significant ( $p<.0001$ ). The overall effect size of the intervention was impressive, approximately .75.

On another metric of program effectiveness, the number of children who scored in the top quartile in national performance was about 40 or 50 percent, This percentage was remarkable given that the project schools were largely poor, minority-serving institutions whose baseline performance on standardized tests was unremarkable to poor. During this time also, longitudinal tracking of students showed a cumulative effect through multiple years' exposure to M+M.

Some teachers became more comfortable and skilled with the approach because they had multiple years of experience.

During this year, M+M activity sequences were altered, again to conform more satisfactorily to students' experience in schools. The initial ordering had a cognitive justification: exercises in ST reasoning were given first to give students a basis for learning math in this modality; instruction in mathematics concepts proper followed. A second change was that the order of games was rearranged to conform to typical teaching sequences. These changes were wrought partly because of teacher feedback. In classroom lessons, teachers wanted to be able to make connections to what students were learning in the M+M games. Moreover, during this time there was some sifting of games: new ones were introduced and others were pulled.

### 2002-2003

By September of 2002, M+M had spread to 16 schools that had both participating ( $n=4,173$ ) and nonparticipating ( $n=1,546$ ) students. The improvement of the math integration component was a major change to M+M. Structurally, math integration was being incorporated into the software rather than being conveyed through a weekly lesson by a teacher. An advantage was greater assurance that each student was getting practice with math integration concepts—in particular, linking the ST representation with more standard language-analytic approaches. The desire was not to circumvent the teacher, but there was a clear realization that students' exposure to the classroom math integration lessons was spotty. Some teachers and schools were dependable in

providing this crucial component, but others did not do so, or taught math integration only irregularly.

In California, the state accountability test -shifted to the CAT 6. Again, the M+M participants exhibited higher overall mathematics learning. On the CAT 6, M+M students had considerably higher average scores ( $M=54.5$ ,  $SD=7.2$ ,  $N=1680$ ) in comparison to nonparticipating students ( $M=30.0$ ,  $SD=10.8$ ,  $N=382$ ). This difference was statistically significant,  $p<.001$ . Considering only classes that completed 40 percent more of the M+M program, the CAT 6 Mathematics mean was higher still ( $M=59.2$ ,  $SD=6.9$ ,  $N=1009$ ). Below that threshold, the achievement distributions were identical to nonparticipating students. Translating this latter advantage into standard deviation equivalents, the overall impact of the program on standardized test scores was approximately two sigma, historically considered the maximum effect achievable by an educational intervention (Bloom, 1984). Similar results favoring the M+M group were found at the third and fourth grade levels (MIND Institute, 2005).

During this year, the California Standards Test became much more significant in state accountability. The number of students who tested at a proficient or higher level was significantly greater for M+M students than for nonparticipants. Differences among groups were more pronounced when participating students were separated into those who completed less than 50 percent of the treatment and those who completed more than 50 percent. Those students who completed less than 50 percent were regarded by the research team as not truly experiencing the intervention. Among full

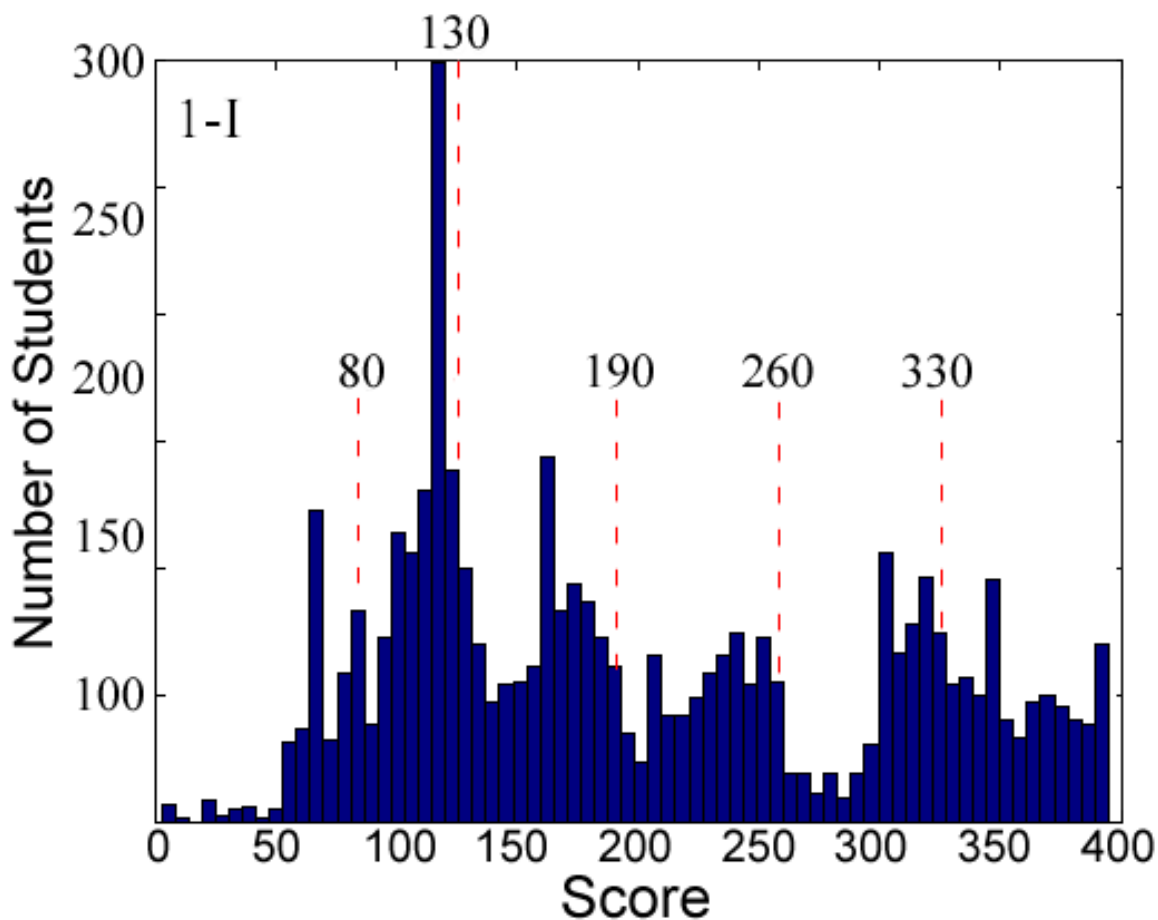
participants in grades 2, 3, and 4, those who completed at least 50 percent of the activities, about 60 percent of students were ranked as being “advanced” or “proficient” (MIND Institute, 2004). Among nonparticipating classes and those students completing less than 50 percent of the activities, only about 40 percent of students were either “advanced” or “proficient.” Also at the 2nd, 3rd, and 4th grade levels, the number of students at “far below basic” was almost nonexistent among participants (MIND Institute, 2004). In other words, among M+M students the entire distribution—the mean and both tails—shifted to higher levels. Intra-school comparisons, which more clearly contrast similar students, also show substantial achievement differences favoring students who engaged in at least 50 percent of M+M activities at grades 2, 3, and 4 (MIND Institute, 2004).

#### 2003-2004

During the 2003-2004 academic year, M+M was implemented in approximately 40 elementary schools. A shrinking percentage of project schools, only nine, had both participating and nonparticipating students. By this time in the project, most schools wanted to implement M+M grade-wide. Presumably, this shift reflected, in part, a greater confidence that M+M offered benefits to all students, and that the approach was workable for most teachers. During this year also, more concepts were covered to reach a progressively wider swath of the accepted mathematics curriculum. The project team understood, especially as state-level accountability systems were rising to full

implementation, that the M+M curriculum needed to span the expected state content to be relevant and acceptable.

In March through May of 2004, individual student records were analyzed to seek patterns of learning progress. Data mining techniques showed that these learning curves clustered into four distinct types. Three of the generic learning curves showed students reaching an impasse at particular points (Hu, Bodner, Jones, Peterson, & Shaw, 2004). Eventually, a culprit was identified. It became apparent that features of the software intended to elicit and maintain student interest were actually having the opposite effect. Some of the software animation, sound effects, background texturing, and extra “fun” activities were undermining students’ focus on the intended content and skills. Rather than co-opting students’ interests to ST learning, the design elements proved to be a distraction. For example, one game incorporated a walkway used by the main character, Jiji. Second by second, the walkway shrunk such that eventually Jiji could not make it across. Interestingly, a histogram showing frequencies of students’ game scores revealed an impasse at 130 points (See Figure 1, below) that corresponded to the shrinking walkway feature. Many students simply could not make it past this rather incidental software feature. Such features were reduced or eliminated.



Feedback to teachers at this point was provided via e-mail and over the web. The curriculum also expanded to include more grade levels. During this year, the fourth grade curriculum was introduced, and work began on K-1 software. School staff—teachers and principals—conveyed the distinct desire to have even younger primary students begin experience with ST concepts. For example, they wanted entering second graders to be proficient in working with proportions.

#### 2004-2005

Preliminary data indicate a performance advantage on mathematics achievement for M+M students. A separation was made post hoc in recognizing

high participants (Y>50) as those who completed at least 50 percent of the prescribed M+M curriculum for a given grade level. Low participants complete less than 50 percent. A third group of nonparticipants (NP) did not experience M+M. On the California Standards Test of mathematics, a higher percentage of high participants were judged “proficient” or “advanced” than was the case among nonparticipants. This pattern held at the three grade levels examined: 2nd grade (Y>50: 54%; NP: 41%), 3rd grade (Y>50: 63%; NP: 40%), and 4th grade (Y>50: 44%; NP: 30%).

In approximately January of 2005, project staff realized that there was an unworkable time lag between feedback cycles from standardized test results to the redesign of software. Results from standardized testing associated with accountability testing were reportable approximately one year after students were assessed. That meant that a full two years were needed from the time of program redesign and implementation to the reporting of results from standardized tests. If changes were made in response to assessment results, these too required time. Yet another year of implementation was followed by a one-year lag as schools waited for test results. Piecing these phases together, the total design/implementation/outcome cycle extended over three years. Of course, these phases were not followed in strict serial fashion; curriculum redesign was ongoing, as was field testing. Still, the true feedback cycle for any particular design plus implementation stretched over several years. The cycle feedback lag impeded software redesign and development.

At this point, and partly in response to these suboptimal development models, the MIND Institute initiated a different development process. Named after the penguin mascot of the Institute, a subdivision called Jiji Labs began to develop new modules on such topics as place value, decimals, and elapsed time. Unit-based pre- and post-tests were modeled after questions on standardized tests. Staff tested the modules rapidly through a succession of pretests, focused games, and post-tests. This allowed field testing to be completed within weeks. The development cycle was thereby compressed from three years to about one month. For example, in teaching place value to first graders, students initially engage in the spatial-temporal problems alone. Pre- and post-tests showed no gains on the standard language-analytic understanding of place value. A short additional experience with the language and notation of place value resulted in significant gains on a second comparable post-test (see Figs. 2 & 3, below).

Fig. 2 Place Value Spatial Game Pre- and Post- Test Accuracy

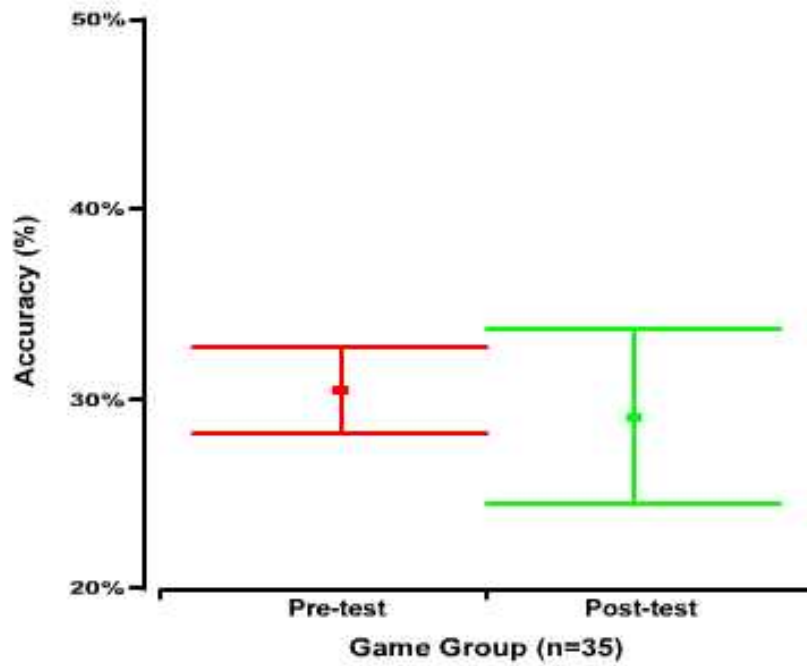
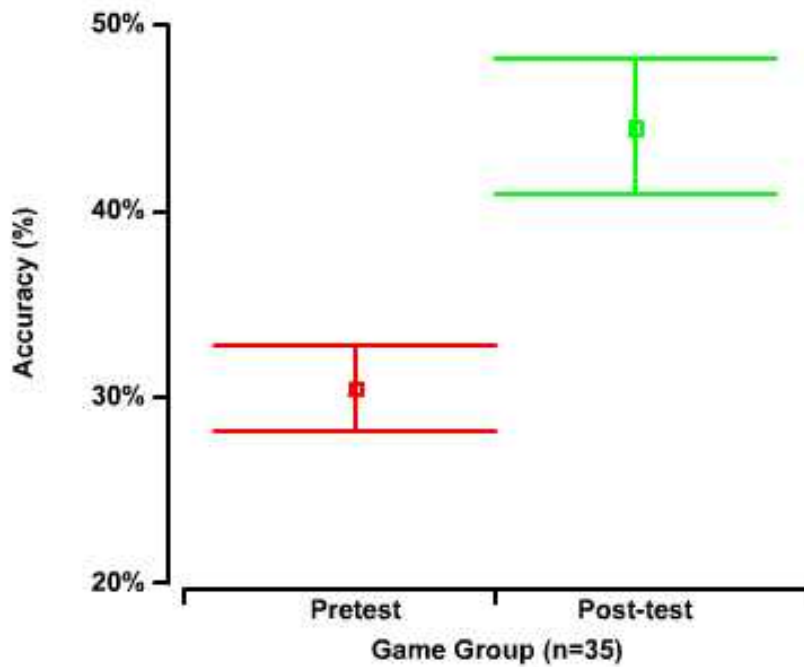


Fig. 3 Place Value Language Integration Pre- and Post- Test Accuracy



During the previous year, the research team noted that students performed differently when taking quizzes online rather than on paper. Kids treated computer-based tests like computer games, sometimes making arbitrary choices or skipping sections. The electronic versions of quizzes were not taken seriously, as if the ephemeral quality of electrons made them trivial in comparison to pencil lead. In response, during this year the project used paper-based tests.

Other changes were made in the design of games. Besides instituting a more rapid development/feedback cycle, the games were redesigned toward greater consistency of format. Previously, each game had its own associated rules. This meant that teachers had to teach rules for each new math topic. To do this, they often had to refer to a manual to acquaint themselves with game rules so that they could provide explanations to the children. This requirement added considerably to the time and effort overhead required of participating teachers. The research team was not at all confident that every teacher in fact introduced the rules of each game; probably some did but others did not.

The redesign of games this year was based on a common and very simple goal—to get Jiji the penguin across the screen from left to right. We might guess that children would quickly become bored with this regularity across games, but so far children seem to enjoy this consistency. The known goal easily bridges one game to another. Rather than relying on external instruction in the game, an interactive demo or tutorial was built into each module to

introduce the basic structure of the game. Another change was to begin each module with an easy task—often trivially easy—to ensure that students were at least initially successful. Previous field testing showed that some students experienced failure at the start and lost confidence.

During this year, feedback to teachers was sent entirely over the internet. Since 2000, the scale of implementation expanded from 4 to 60 schools. Also, the number of implemented games increased. Ultimately, the MIND Institute server could not handle the data volume and the necessity of interfacing with several remote database applications. Data gathered from schools were often corrupted. This prompted a switch to an industry-standard server architecture. The result was an efficient data gathering process, ending a long-term struggle to collect complete and timely data from participating schools. Now, even with 13,000 participating students, data could be collected. In turn, classroom-level data could be analyzed, aggregated, summarized, and plotted—and then returned to teachers real-time via a web portal to guide their instruction planning decisions.

### 2005-2006

Starting in the fall of 2005, the curriculum will expand to include 98 percent of the math standards. Consequently, the M+M curriculum will comprise a larger percentage of content assessed on standardized tests. In other ways, too, the software intervention is undergoing a very substantial revision. Improvements to some specific games have been substantial. In addition, language integration will expand dramatically to include more than half

of the software. Feedback systems, from schools and back to schools, will be real-time.

Another program change was a response to feedback from teachers. Many teachers wanted quizzes to be incorporated into the software. These quizzes were constructed and incorporated into M+M. One problem with this system was that the quizzes did not provide reliable information about students' learning. Some students treated quizzes in a fashion similar to the low-stakes games. They felt free to experiment, skip screens, and even try making errors intentionally to see what would happen. In other words, the mindset that functioned well in the regular M+M activities did not transfer effectively to the online quizzes. These will be eliminated during the 2005-2006 school year.

## Discussion

### M+M Effects

Our data show that the M+M intervention produced substantial gains in mathematics achievement among participating second graders in comparison to control group students. During the school years for which we have robust data, M+M students displayed a mathematics achievement advantage of at least 14 percentile points over control students. That advantage grew each year even as the intervention spread to larger numbers of sites. The effects were manifest on standardized measures of broad mathematics achievement, the Stanford 9 and the CAT 6. Advantages were also evident in achievement of proficiency on the California Mathematics Standards.

Our data affirm that a spatial-temporal approach to learning mathematical concepts, allied with music instruction, can produce gains in proficiency with mathematics concepts and skills among children. Music instruction can be allied profitably with spatial-temporal mathematics instruction to produce increases in broad mathematics learning. Even so, it is hard to parcel program effects on mathematics achievement precisely. Much of the benefit of M+M appears not to be directly attributable to the music component. Instead, most of the boost in mathematics learning is attributable to the spatial-temporal approach of teaching mathematics concepts. We estimate that about 80 percent of the M+M effect is associated with our adoption of a spatial-temporal approach, and the remaining 20 percent can be linked to the music component. Admittedly, these are rough estimates and are somewhat speculative. Relatedly, we do not yet have a definitive answer to the question: Would there be an enhancement to mathematics achievement based on music instruction alone?

The cumulative findings of this multi-year research project imply that a large segment of students, perhaps most, could benefit from an approach to learning mathematics that appropriates ST reasoning along with music training. The use of ST reasoning and representations might hold special promise with English language learners because of its relative de-emphasis of language—specifically, mathematical terms expressed in English.

#### The Design Experiment Approach

The research project presented here illustrates how an educational model can co-evolve with an educational product. This is the core logic of a design

experiment: New stages in the evolution of an educational intervention are shaped by data from previous stages. Still, there are aspects of our experience that are not always considered in the literature on design experiments. Our experience with M+M permits us to make the following observations on this multi-year, iterative project:

- Precise and reliable measurement using formal instruments functioned largely to tell interested parties that the basic intervention was effective. Achievement tests assured teachers, school administrators, parents, sponsors, and the research team that our program was working, and therefore worthwhile. Their function was significantly motivational, and partly political, in that good results from tests sustained the will to continue. Traditional instrumentation was less important in informing decisions on design alterations to the intervention.
- Design alterations to the M+M intervention were sometimes informed by more detailed examination of performance on home-grown instruments (which led, for example, to a realization that students also needed a language integration component) and by tracking student progress through the software (on which data mining techniques led eventually to the identification of impasses in learning curves).
- Data that informed design decisions were often informal, contextual, and personal. Teachers often had specific ideas about which program functions worked well and which needed alteration. This told us that even when reliable instrumentation is included in the experiment design, it can

be other, more ideographic and detailed sources of information that have practical influence on program design decisions. We suspect that this pattern is typical of other multi-year, iterative projects. Whether it is inevitable or not is unclear.

- In the course of this project's evolution, there was a drift away from controlled research designs with more-or-less clear demarcations between treatment and control groups. This drift coincided with the expansion of the project to larger numbers of participating schools and with a growing confidence that the intervention was essentially effective. Motivational and political functions served by data from standardized tests became less important. As before, design improvements were often motivated by other, less formal information sources. Inevitably, this weakened inferences from data to generalizable conclusions, but did not slow down the development cycle, or dampen the effectiveness or expansion of the program. These changes might be natural or typical concomitants of scaling up.

In aggregate, the M+M project demonstrates the viability of the design experiment approach to educational interventions for advancing student learning and the theories on which effective interventions are based. In our experience, though, the program drifted from the rigorous design ideals of pure experiments. Reliable, standardized instrumentation became less important over time. This is partly because more formal designs and instruments functioned initially to convince researchers, teachers, parents, and others that

the M+M program was worthwhile. More specific design decisions were instead informed by detailed analysis of home-grown (and less reliable) instrumentation, data from performance on the computer-based games, and feedback from teachers. All things considered, the design experiment logic, with some variation, proved to be vital to the ongoing improvement of the design, scope, and effectiveness of our project.

## References

- Bloom, B. S. (1984). The 2 sigma problem: The search for methods of group instruction as effective as one-on-one tutoring. Educational Researcher, 13(6), 4-16.
- Bodner, M., Kroger, J., & Fuster, J.M. (1996). Auditory memory cells in dorsolateral prefrontal cortex. Neuroreport, 7, 1905-1908.
- Bodner, M., Muftuler, L., Nalcioglu, O., & Shaw, G.L. (2001). FMRI study relevant to the Mozart effect: Areas involved in spatial-temporal reasoning. Neurol. Research, 23, 683-690.
- Bodner, M., Shaw, G.L. (2001). Symmetry operations in the brain: Music and reasoning. In Algebraic Methods in Physics, edited by Y. Saint-Aubin and L. Vinet. Springer NY, pp. 17-35.
- Brown, A., L. (1992). Design experiments: Theoretical and methodological challenges in creating complex interventions in classroom settings. The Journal of the Learning Sciences, 2(2), 14-178.
- Cobb, P., Confrey, J., diSessa, A., Lehrer, R., & Schauble, L. (2003). Design experiments in educational researcher. Educational Researcher, 32(1), 9-13.
- Cobb, P., & Gravemeijer, K. (this volume). Outline of a method for design research in mathematics education.
- Fuster, J.M., Bodner, M., & Kroger, J. (2000). Cross-modal and cross-temporal association in neurons of frontal cortex. Nature, 405, 347-351.

- Graziano, A. B., Peterson, M., & Shaw, G. L. (1999). Enhanced learning of proportional math through music training and spatial-temporal training. Neurological Research, 21, 139-152.
- Hetland, L. (2000). Listening to music enhances spatial-temporal reasoning: Evidence for the "Mozart Effect." Journal of Aesthetic Education, 34(3,4), 105-148.
- Hu, W., Bodner, M., Jones, E. G., Peterson, M. R., & Shaw, G. L. (1994, May). Dynamics of innate spatial-temporal learning process: Data driven education results identify universal barriers to learning. Paper presented at the International Conference on Complex Systems.
- Hughes, J.R., Daaboul, Y., Fino, J.J., & Shaw, G.L., (1998). The "Mozart Effect" in epileptiform activity. Clin. Electroencephalograph 29, 109-119.
- Hughes, J.R., Fino, J.J., Melyn, M.A. (1999). Is there a chronic change of the "Mozart Effect" on epileptiform activity? A Case Study. Clin. Electroencephalography, 30, 44-45.
- Huttenlocher, P.R. (2002). Neural plasticity: The effects of environment on the development of the cerebral cortex. Cambridge, MA: Harvard University Press.
- Leng, X., & Shaw, G.L. (1991). Toward a neural theory of higher brain function using music as a window. Concepts in Neuroscience, 2, 229-258.
- McGrann, J.V., Shaw, G.L., Shenoy, K.V., Leng, X., & Mathews R.B. (1994). Computation by symmetry operations in a structured model of the brain. Physical Review, 49, 5830-5839.

MIND Institute. (2004). STAR treatment effect on mathematics performance levels of 2nd, 3rd, and 4th grade students measured using the California Advanced Test Form 6 and the California Standards Test, 2002/3.

Unpublished research bulletin available at [www.mindinstitute.net](http://www.mindinstitute.net).

Mountcastle, V.B. (1978). An organizing principle for cerebral function: the unit module and the distributed system. In G.M. Edelman & V.B. Mountcastle (Eds.), The mindful brain (pp. 1-50). Cambridge, MA: MIT Press.

Mountcastle, V.B. (1997). The columnar organization of the neocortex. Brain, 120, 701-722

Muftuler, T., Bodner, M., Shaw, G.L., & Nalcioglu, O., (2004). fMRI study to investigate spatial correlates of music listening and spatial-temporal reasoning. Abstr. Annual meeting of the Int. Soc. of Mag. Resonance in Medicine, 12th annual meeting.

Rauscher, F.H., Shaw, G.L., & Ky, K.N. (1993). Music and spatial task performance. Nature, 365, 611.

Rauscher, F.H., Shaw, G.L., & Ky K.N. (1995). Listening to Mozart enhances spatial-temporal reasoning: towards a neurophysiological basis. Neuroscience Letters, 185, 44-47.

Rauscher, F.H., Shaw, G.L., Levine, L.J., Wright, E.L., Dennis, W.R., & Newcomb, R.L. (1997). Music training causes long-term enhancement of preschool children's reasoning. Neurological Research, 19, 2-8.

Rauscher, F.H., & Shaw, G.L. (1998). Key components of the Mozart Effect. Perceptual and Motor Skills, 86, 835-841.

- Rauscher, F.H., Robinson, K.D., & Jens, J.J. (1998). Improved maze learning through early music exposure in rats. Neurological Research, *20*, 427-432.
- Sarnthein, J., von Stein, A., Rappelsberger, P., Petsche, H., Rauscher, F.H., & Shaw, G.L. (1997). Persistent patterns of brain activity: An EEG coherence study of the positive effect of music on spatial-temporal reasoning. Neurological Research, *19*, 107-116.
- Schellenberg, E.G. (2004). Music lessons enhance IQ. Psychological Science, *15*, 511-514.
- Shaw, G.L. (2000). Keeping Mozart in mind. San Diego: Academic Press.
- Shenoy, K.V., Kaufman, J., McGrann, J.V., & Shaw, G.L. (1993). Learning by selection in the trion model of cortical organization. Cerebral Cortex, *3*, 239-248.