Research Report

Supplementing a Traditional Math Curriculum With an Inquiry-Based Program: A Pilot of Math Out of the Box

JoAnn L. Rock

Rosalea Courtney

Philip G. Handwerk

May 2009

ETS RR-09-17

ETS

Listening. Learning. Leading.®

Supplementing a Traditional Math Curriculum With an Inquiry-Based Program: A Pilot of Math Out of the Box

JoAnn L. Rock and Rosalea Courtney
ETS, Princeton, New Jersey

Philip G. Handwerk
Law School Admission Council, Newtown, Pennsylvania

May 2009

As part of its nonprofit mission, ETS conducts and disseminates the results of research to advance quality and equity in education and assessment for the benefit of ETS's constituents and the field.

To obtain a PDF or a print copy of a report, please visit:

http://www.ets.org/research/contact.html

Copyright © 2009 by Educational Testing Service. All rights reserved.

ETS, the ETS logo, and LISTENING. LEARNING. LEADING. are registered trademarks of Educational Testing Service (ETS).

Abstract

This study examined math achievement of elementary school students when Math Out of the Box (MTB), an inquiry-based math program, was used to supplement curriculum. The sample consisted of 767 New Jersey students in the third, fourth, and fifth grades, with approximately one third using MTB. Math achievement was measured by an assessment developed by ETS and by New Jersey's standardized test of math proficiency (NJ ASK). On the ETS assessment, a small, statistically significant difference was found in each of the three grades between students who used MTB and those who did not. On NJ ASK, a small, statistically significant difference was found in the third grade only. While these findings are an important step in establishing the efficacy of MTB, selection bias may weaken the causal inferences drawn.

Key words: Math proficiency, Math Out of the Box (MTB), NJ ASK, ANCOVA, math education, inquiry-based learning

Acknowledgments

This evaluation required the collaboration of many stakeholders. This was a positive experience largely due to the creative and skillful efforts of Sona Polakowski. A number of ETS staff members were major contributors to this evaluation. This included assessment specialists Jeff Haberstroh and Michael Renz, who were responsible for the development of the pre- and post-year assessments, and Eleanor Horne, who participated in stakeholder meetings and hosted many of the MTB professional development sessions. We also acknowledge Kathleen Howell, who provided administrative support.

Introduction

The purpose of this study was to examine the effect of an inquiry-based mathematics program on student math achievement. Math Out of the Box (MTB) is a K-5 mathematics program developed by the College of Engineering and Science at Clemson University. MTB uses active learning activities to engage students in inquiry-based learning, with an emphasis on reasoning, problem-solving, and higher order thinking skills. The program includes four K-5 strands—*Developing Algebraic Thinking, Developing Geometric Logic, Developing Measurement Benchmarks,* and *Developing Number Concepts.* The strands can be used independently or collectively.

In this study, the *Developing Algebraic Thinking* and *Developing Geometric Logic* strands of MTB were used as a supplement to the more traditional math curriculum offered in a suburban New Jersey school district. Within each strand, a grade level module included inquiry-based curriculum materials and instructional resources for a class of students. The modules are marketed by Carolina Biological Supply Company and include a large box of materials sufficient for approximately 30 students, along with the *MTB Teacher Guide*. The box includes a variety of materials, such as sets of colored blocks representing different three-dimensional shapes, white boards, calculators, and protractors. The *MTB Teacher Guide* includes quizzes and tests for the module and reproducible practice pages.

MTB uses an inquiry approach to math instruction. This type of approach differs from the traditional approach in three important ways: the role of the teacher, the role of the student, and the hypothesized process through which knowledge is acquired. Regarding the first difference, in the traditional approach to teaching math, the teacher is viewed as controlling the learning process through the dispensing of information (Whitehurst, 2003). This is juxtaposed with the inquiry-based approach, where the role of the teacher is viewed as a facilitator of students' individual acquisition and construction of knowledge (Cobb & Bauersfeld, 1995; Kaser & Bourexis, 1999). In other words, the teacher's role is to guide students in developing their own understanding of mathematical principles.

The second difference is in the role of the student in the two approaches. The inquiry approach places the student at the core of the learning process, constructing new knowledge through interaction with guided materials and activities (Grouws & Cebulla, 2000; National Council of Teachers of Mathematics [NCTM], 2000). On the other hand, the traditional approach

is more textbook driven and views the student as a more passive receiver of information (Sood & Jitendra, 2007). The third difference involves the process through which knowledge is hypothesized to be acquired in the two approaches. The traditional approach tends to focus on mathematic procedures developed sequentially along a hierarchy of skills, while the inquiry approach views learning as a dynamic acquisition of concepts and skills, which can be applied across areas of mathematics and other disciplines (Hope, 2007; Schoor & Amit, 2005).

In the study reported here, the effect of MTB on student achievement in third, fourth, and fifth grade was estimated using a quasi-experimental research design. In this design, the math achievement of students in classes that used MTB (MTB group) as a supplement over the 2006-2007 school year was compared to that of students in classes that did not (non-MTB group). The student sample consisted of 767 students across the three grades. Approximately one third of the students were in classes that used MTB, and the others were not. The effect of MTB was defined as the average difference in math achievement between the two groups. The analysis was conducted within grades. In order to provide some degree of statistical control on the estimates, analysis of covariance (ANCOVA) was used to adjust the end-of-year differences in math scores for observable baseline characteristics related to math achievement. Two math outcomes were used to measure math achievement—an assessment developed by ETS and New Jersey's standardized math proficiency test (NJ ASK).

While quasi-experimental designs are important first steps in establishing the efficacy of a program, selection bias may weaken the causal inferences that can be drawn. In other words, the characteristics of the teachers and students in the MTB and non-MTB groups may differ in ways that influence the observed differences in outcomes at the end of the study period. When random assignment is used to assign teachers and students to treatment and comparison groups, rather than a nonequivalent group design such as used here, the causal inferences that can be drawn are stronger.

Method

Procedures

This study was conducted in five elementary schools in a New Jersey suburban school district over the 2006-2007 school year. All 52 teachers in the third, fourth, and fifth grades participated in the study. Twelve teachers from the district were trained in the use of MTB. (See Table 1.) The teachers attended 3-day workshops given by MTB facilitators, who used

professional development modules developed for MTB. The 12 teachers implemented the program in their classes over the 2006-2007 school year as a supplement to the district's current math curriculum. In addition to the training, the teachers were given the required kits needed to implement the program in their classes over the study period. Over the same time period, the remaining 40 teachers utilized the current district math curriculum.

It is important to note that the teachers in the MTB group volunteered to use the program in their classes. Since the teachers were not randomly assigned to use the program, the two groups of teachers may differ in ways that influenced their students' math achievement—other than the use of MTB. Therefore, the inference that MTB was the cause of observed differences between the two groups of students at the end of the study period is weaker than it would be if random assignment were used to assign the teachers.

Table 1

Math Out of the Box (MTB) and Non-MTB Teachers Across Grades

Grade	Total	MTB	Non-MTB
Grade 3	17	6	11
Grade 4	16	4	12
Grade 4/5	3	0	3
Grade 5	16	2	14
Total	52	12	40

Training

The teachers in the MTB group were trained in the use of the MTB modules *Developing Algebraic Thinking* and *Developing Geometric Logic* in two 3-day professional development workshops. A facilitator module was included with each grade level module of each curriculum strand. The teachers were trained in two cohorts with an experienced facilitator, who introduced the scope of each strand. In the workshops, the teachers were provided with hands-on experiences within grade levels and with opportunities to become familiar with the teacher's guide and student manipulatives. Because the school district was piloting the curriculum, feedback about the curriculum was solicited from participants during follow-up sessions.

While the study year was the 2006–2007 school year, the training and implementation of MTB in the district began in 2005. The first group of teachers (Cohort 1) participated in professional development in the spring of 2005. The Cohort 1 teachers began teaching the *Developing Algebraic Thinking* and *Developing Geometric Logic* MTB strands during the year. In February 2006, the second group of teachers (Cohort 2) was introduced to MTB and began teaching the algebra strand and the geometry strand in the spring of 2007. Therefore, in the 2006–2007 study year, the 12 teachers who implemented MTB had different years of experience teaching the modules, depending on their cohort assignment.

Implementation of the Treatment

An important first step in estimating the effect of an intervention is ensuring that the program was implemented according to its established principles. During the study period (Cook & Campbell, 1979; Fitzpatrick, Sanders, & Worthen, 2004). To examine the implementation of MTB in the school district over the 2006-2007 school year, algebra lessons in 10 classrooms were observed. It is important to note that the teachers in the observation sample volunteered to be observed, and this may have biased the observations, since teachers who were more comfortable with the program may have volunteered. While these observations do not provide a complete picture of the implementation, they do provide a rudimentary measure of the effectiveness of implementing the program in the classrooms.

In general, the teachers who were observed were effective in implementing the components of MTB. In most observations, students were actively engaged in math activities throughout the lesson. Students were often given opportunities to work together in small groups in a collaborative way. The small group work required students to solve problems together or to construct lists of attributes. These student-student conversations appeared to be moving them towards greater understanding or comfort with the mathematical concepts. In most lessons, the mathematics was standards based, appropriate, and challenging. Typically, students had opportunities to communicate their understanding through discussion and/or writing, and in some instances through both modalities. The design of the lessons allowed students across all the grades to apply their understanding in activities and problems that went beyond drill and practice.

Measures

The pre- and post-year ETS assessments for each grade consisted of 18 multiple-choice questions and three constructed-response questions. Two scores were constructed for each pre- and post-year assessment per grade. The total score consisted of two components, a multiple-choice section and a constructed-response section. The highest possible total score for third graders was 25, and for fourth and fifth graders it was 27. The highest constructed-response score was 7 for third grade and 9 for the fourth and fifth grades. The constructed-response items were scored using rubrics that yielded from two to four points maximum, depending on the item.

The ETS pre- and post-year assessments were designed to cover the content of the MTB strands. Forms were assembled to content specifications, making sure that each major content area was sufficiently represented in both forms (pre- and post-year). The pre- and post-year tests were designed to be equivalent forms. Out of the items used, 9 of 21 used in the pre-year assessments at each grade level appeared again in the post-year assessments. The additional items in the post-year assessment used different problems that addressed the same content areas as those in the pre-year test.

For constructed-response items, the assessment staff developed 2- to 4-point rubrics, depending on the nature of the question. Teachers were recruited from nearby districts to score the constructed-responses in a 1-day session. Training was provided in the use of the rubrics for each of the three questions scored at each grade level.

The second outcome measure was New Jersey's standardized math proficiency test (NJ ASK), which was employed for all three grades as a post-year measure. NJ ASK is a state assessment of student achievement in language arts, math, and science that is taken by all New Jersey third, fourth, and fifth graders. The test assesses student achievement in the knowledge and skills defined by the New Jersey Core Curriculum in language arts, literacy, mathematics, and science. The test consists of two major types of questions, multiple-choice and open-ended questions. The topic areas covered in the math section of the test are numbers, numerical operations, geometry and measurement, patterns and algebra, data analysis, probability, and discrete mathematics. The scores used as outcome measures in the analysis were the total raw score and the number operations, math patterns, and problem solving scores for each grade.

Sample

The study sample consisted of all third, fourth, and fifth grade students in a suburban New Jersey school district. The analytic sample consisted of students across the three grades that had completed pre- and post-year assessments. This sample consisted of 767 students; 265 were from the third grade, 242 from the fourth grade, and 260 from the fifth grade. About half of the sample were White; 20%, Black; 10%, Hispanic; and 15%, Asian American. Approximately 7% of the students were identified as limited English proficient (LEP), and 15% of the children were from families classified as low socio-economic status (SES). The demographic distributions were similar across the three grades, with slight variations; see Table 2.

Table 2

Student Demographics at Each Grade Level (Numbers and Percentages)

Demographics	Grade level					
	Third		Fourth		Fifth	
	N	%	*N*	%	*N*	%
White	144	54.34	134	55.37	143	55.00
Black	50	18.87	44	18.18	56	21.54
Hispanic	23	8.68	27	11.16	28	10.77
Asian American	48	18.11	37	15.29	33	12.69
LEP	9	3.40	2	0.83	10	3.85
Low SES	33	12.45	41	16.94	44	16.92
N	265		242		260	

Note. LEP = limited English proficient, SES = socio-economic status.

Differences Between Treatment and Comparison Groups at Baseline

In this design, the effect of MTB is estimated as the average difference in math achievement between students in the MTB group and those in the non-MTB group at the end of the school year for each of the grades. Two scores were separately used to measure math achievement: scores on the ETS post-year assessment and the scores on the NJ ASK. The validity of the estimate depends upon the two groups of students being similar enough at baseline on characteristics related to math achievement so that an assumption can be made that the

observed differences in math achievement at the end of the period are due to MTB, rather than to differences in the characteristics of students. This is particularly important, since the two groups of students were not randomly assigned to the treatment and comparison groups and may differ in characteristics correlated to math achievement. To examine this issue, the baseline demographics and test scores of the MTB and non-MTB groups of students were analyzed to see if the groups differed in discernible ways. These differences were analyzed within grade. See Table 3.

As shown in Table 3, in the third and fifth grades, the demographic characteristics of the students in the MTB group and those in the non-MTB group were similar. There were no statistically significant differences between the two groups. For example, in both grades the ethnicity of the students was about the same.

Table 3

Demographics of Sample Overall and by Math Out of the Box (MTB) Status

Characteristics	Full		MTB		Non-MTB		Chi-sq
	N	%	*N*	%	*N*	%	*p*-value
Third grade							
Ethnicity							
White	144	54.34	64	59.81	80	50.63	.43
Black	50	18.87	16	14.95	34	21.52	
Hispanic	23	8.68	8	7.48	15	9.49	
Asian American	48	18.11	19	17.76	29	18.35	
SES							
Low	232	87.55	11	10.28	22	13.92	.38
Other	33	12.45	96	89.72	136	86.08	
LEP							
Yes	9	3.40	1	.93	8	5.06	.07
No	256	96.60	106	99.07	150	94.94	

(Table continues)

Table 3 (continued)

Characteristics	Full		MTB		Non-MTB		Chi-sq
	N	%	*N*	%	*N*	%	*p*-value
Fourth grade							
Ethnicity							
White	134	55.37	47	58.02	87	54.04	.00
Black	44	18.18	10	12.35	34	21.12	
Hispanic	27	11.16	4	4.94	23	14.29	
Asian American	37	15.29	20	24.69	17	10.56	
SES							
Low	41	16.94	5	6.17	36	22.36	.00
Other	201	83.06	76	93.83	125	77.64	
LEP							
Yes	2	0.83	0	0	2	1.24	.31
No	240	99.17	81	100.00	159	98.76	
Fifth grade							
Ethnicity							
White	143	55.00	17	48.57	126	56.00	.58
Black	56	21.54	7	20.00	49	21.78	
Hispanic	28	10.77	6	17.14	22	9.78	
Asian American	33	12.69	5	14.29	28	12.44	
SES							
Low	44	16.92	7	20.00	37	16.44	.60
Other	216	83.08	28	80.00	188	83.56	
LEP							
Yes	10	3.85	4	11.43	6	2.67	.01
No	250	96.15	31	88.57	219	97.33	

Note. LEP = limited English proficient, SES = socio-economic status.

Since previous math achievement is highly correlated with future math achievement, differences in baseline test scores between groups were also examined; see Table 4. The baseline math scores of the groups did not differ significantly in the third and fifth grades. The average total score on the ETS assessment was 8.83 for the MTB group and 9.02 for the non-MTB group. In the fifth grade, the average math score in the MTB group was 10.63 and the average math score in the non-MTB group was 11.36. In the fourth grade, however, there were statistically significant differences between the MTB and non-MTB groups in demographic characteristics and math scores.

Table 4

Baseline Math Scores Overall and Math Out of the Box (MTB) Status

Group	Type of score								
	Total			Multiple-choice			Constructed-response		
	Mean	*SD*	*p*-value	Mean	*SD*	*p*-value	Mean	*SD*	*p*-value
				Third grade					
MTB	8.83	3.78	.675	7.01	3.04	.891	1.82	1.30	.207
Non-MTB	9.02	3.62		6.96	2.57		2.06	1.65	
				Fourth grade					
MTB	15.18	4.59	.000	10.59	3.17	.000	4.59	2.05	.000
Non-MTB	11.42	3.85		8.29	2.61		3.13	1.89	
				Fifth grade					
MTB	10.63	4.15	.413	7.34	2.72	.289	3.29	2.01	.756
Non-MTB	11.36	5.06		7.95	3.21		3.41	2.29	

In the fourth grade, there were more Black and Hispanic students and fewer Asian American students in the non-MTB classes than in the MTB classes. In addition, there were statistically significant differences between the average student math scores in the MTB classes

(15.8) and the non-MTB classes (11.42). Furthermore, the average baseline test scores of the White, Asian American, and Hispanic students in the non-MTB classes in fourth grade were lower than those of students in the MTB groups. These data are displayed in Table A2 in the appendix. These factors should be taken in account when examining the results.

Analytic Strategy

To improve the accuracy of the estimate, analysis of covariance (ANCOVA) was used to adjust the end-of-year differences in math scores for observable baseline characteristics related to math achievement. The effect of MTB was defined as the adjusted average difference in math achievement between students in the MTB and non-MTB groups at the end of the study period for each grade. The average treatment group-comparison group differences were adjusted to minimize observed baseline differences between the groups that may have influenced the observed differences at the end of the study period. Without adjusting for these differences, the accuracy of the evaluation of the program effect would be weakened, since the comparison group used to isolate the treatment effect would be notably different in ways that would likely be correlated with the outcome measures (Cook & Campbell, 1979). Therefore, it would not be possible to conclude with reasonable confidence that the difference was attributed to MTB.

Three baseline measures were included in the statistical model, ethnicity (which was measured as a series of indicator variables), gender, and baseline ETS pretreatment math score. An initial analysis was conducted to identify baseline measures to include in the statistical model. Variables were included that differed between treatment and comparison groups at baseline and were correlated to the outcome measures.

Equation 1 represents the regression model used to statistically adjust the treatment-comparison difference and to create adjusted means for each group.

$$Y = \beta_0 + \beta_1(X_1) + \sum_{i=2}^{6} \beta_i(X_i) + E \qquad (1)$$

Let Y = end of year math score, β_1 = treatment effect, β_2 = covariate effect of the baseline math score, $\beta_{3...5}$ = covariate effect of ethnicity (Black, Hispanic, and Asian American indicators), β_6 = covariate effect of gender, X_1 = treatment indicator, X_2 = baseline math

score, $X_{3..5}$ = ethnicity indicators (Black, Hispanic, and Asian American), and X_6 = gender indicator. In this equation, β_1 represents the adjusted effect of MTB on student math achievement. Statistical significance was set at .05 for a one-tailed test.

The same regression model was used to adjust the means for the treatment and comparison groups. Adjusted means, or least squares means, are predicted population margins or estimates of the marginal means over a balanced population (Kleinbaum, Kupper, & Muller, 1988). The means are adjusted for differences in covariates at baseline by artificially assuming that treatment and comparison groups have the same set of mean covariates and distributions (Rossi, Lipsey, & Freeman, 2004).

Effects were estimated separately for third, fourth, and fifth grade. Within each grade, several outcome measures were used. Effects were estimated on the total score from the ETS assessment and separately on the multiple-choice and constructed-response subscores. Four scores from NJ ASK were used as outcome measures: the total raw score and the scores for number operations, math patterns, and problem solving. The models used for different outcome measures were the same, except the baseline math score that was included differed depending on the outcome measure used. For example, when estimating the effect of MTB on the total score, the baseline total score was used.

Results

When the outcome measures were statistically adjusted for observed baseline differences on available demographics and math scores, they indicated that students who used MTB as a supplementary curriculum did somewhat better on the ETS assessment at the end of the year than students who did not use MTB. This was true across all three grades. While the differences between the groups were statistically significant, the effect was small. The results are summarized in Table 5.

Third graders who used MTB had an average adjusted total score of 17.53 on the ETS assessment at the end of the year, compared to an average score of 16.13 for students who did not use MTB. This difference of 1.40 was statistically significant.

While the reported differences are statistically significant, effect sizes provide an indicator of the magnitude of the differences and allow for comparisons across grades. The effect size in the third grade for the total math score is .33. According to conventional standards, an

effect size between .20 and .49 is considered to be small, an effect size between .50 and .79 is considered to be medium, and an effect size of .80 or more is considered to be large. Thus, the effect size of .33 is considered to indicate a small effect. Effect sizes are summarized in Table 6.

Table 5

Effect of Math Out of the Box (MTB) on Math Achievement: ETS Assessment

Type	Adjusted means			
	MTB	Non-MTB	Diff	*p*-value
		Grade 3		
Total	17.53	16.13	1.40	.001
Multiple-choice	12.51	11.72	0.79	.010
Constructed-response	4.98	4.44	0.55	.016
		Grade 4		
Total	15.04	13.12	1.92	.000
Multiple-choice	11.83	9.98	1.85	.000
Constructed-response	3.46	3.02	0.44	.074
		Grade 5		
Total	17.59	14.57	3.02	.000
Multiple-choice	12.35	10.46	1.88	.000
Constructed-response	5.21	4.11	1.10	.006

The fourth graders who used MTB also did somewhat better. The average adjusted total score of students in MTB was 15.04 compared to an average score of 13.12 for students who did not use MTB. This is a difference of 1.92, and the effect size is .41. The fifth grade adjusted average score for students in MTB was 17.59, compared to 14.57 for the non-MTB students. The difference is 3.02 points, and the effect size is .51. (It is important to note when considering the fifth grade findings that the number of students who used MTB in the fifth grade was very small—only 35 students.)

Looking at the multiple-choice and constructed-response scores across the grades, the effect of MTB was relatively consistent across these two components of the assessment. The exception was the fourth grade, where most of the difference between the two groups appeared

on the multiple-choice section. (See Table A7 in the appendix for a summary of unadjusted and adjusted means.)

Table 6

Effect of Math Out of the Box (MTB) on ETS Math Assessment Across Grades

	Third	Fourth	Fifth
		Total	
MTB	17.53	15.04	17.59
Non-MTB	16.13	13.12	14.57
Difference	1.40	1.92	3.02
Effect size	.33	.41	.51
		Multiple-choice	
MTB	12.51	11.83	12.35
Non-MTB	11.72	9.98	10.46
Difference	0.79	1.85	1.88
Effect size	.29	.54	.50
		Constructed-response	
MTB	4.98	3.46	5.21
Non-MTB	4.44	3.02	4.11
Difference	0.54	0.44	1.10
Effect size	.36	.24	.40

NJ ASK

The effect of MTB was analyzed for the total NJ ASK math raw score and three subtests: number operations, math patterns, and problem solving (see Table 7). There was a small but statistically significant effect of MTB on NJ ASK adjusted scores in the third grade. There were no effects for the fourth and fifth grade scores. (The unadjusted means are summarized in Table A8 in the appendix.)

In third grade, there was a small statistically significant effect on the total NJ ASK raw score and on math patterns when the differences were adjusted for baseline demographic and math achievement differences between the groups. The students' pre-year scores on the ETS

math assessment were used to adjust the NJ ASK scores for baseline treatment group-comparison group differences in math achievement. The effect size on the raw total score was .17. The effect size on math patterns was .21. There was a small statistically significant difference on the math pattern subscore in the third grade. The average adjusted math pattern score in the MTB classes was 5.86, and in the non-MTB classes the average adjusted score was 5.48. There were no statistically significant differences on the number operations or problem solving subscores. The average number operations subscore in the MTB classes was 5.86, and in the non-MTB classes the average score was 5.48. The problem solving subscores in the MTB classes was 7.85, and in the non-MTB classes the score was 7.29.

Table 7

Effect of Math Out of the Box (MTB) on NJ ASK

Type	Adjusted means			
	MTB	Non-MTB	Diff	*p*-value
Grade 3				
Total (raw)	23.35	22.27	1.08	.04
Number operations	6.49	6.37	0.12	.47
Math patterns	5.86	5.48	0.38	.05
Problem solving	7.85	7.29	0.56	.06
Grade 4				
Total (raw)	26.42	24.90	1.52	.07
Number operations	9.25	8.78	0.47	.14
Math patterns	6.40	6.20	0.20	.48
Problem solving	14.09	13.19	0.90	.09
Grade 5				
Total (raw)	25.82	24.89	0.93	.24
Number operations	7.10	7.03	0.07	.83
Math patterns	6.85	6.60	0.25	.42
Problem solving	15.10	15.03	0.07	.90

In the fourth and fifth grades, there were no statistically significant treatment-comparison differences on the adjusted NJ ASK total scores or subscores. In the fourth grade, the average adjusted NJ ASK total raw score in the MTB classes was 26.42, and in the non-MTB classes the average score was 24.90. The NJ ASK adjusted subscore in number operations was 9.25 in the MTB classes and 8.78 in the non-MTB classes. The NJ ASK subscore in math patterns was 6.40 in the MTB classes and 6.20 in non-MTB classes, and the problem solving scores were 14.09 and 13.19 respectively. In the fifth grade, the total raw score in MTB classes was 25.82 and in non-MTB classes the score was 24.89. There were no discernible treatment-comparison differences in the number operations subscores, the math patterns subscores, or the problem solving subscores. The MTB number operations score was 7.10, compared to 7.03 in the non-MTB class. The math patterns average adjusted MTB subscore was 6.85, and the non-MTB score was 6.60. The problem solving subscore in the MTB classes was 15.10, and that for the non-MTB classes was15.03. See Tables A4–A6 for a summary of NJ ASK scores for the full sample and for subgroups stratified by demographics.

Although initial findings about the impact of MTB are promising, a large-scale evaluation study employing random assignment is an appropriate next step.

References

Cobb, P., & Bauersfeld, H. (Eds.). (1995). *The emergence of mathematical meaning*: *Interaction in classroom cultures*. Hillsdale, NJ: Erlbaum.

Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation: Design & analysis for field settings*. Boston: Houghton Mifflin.

Fitzpatrick, J. L., Sanders, J. R., & Worthen, B. R. (2004). *Program evaluation: Alternative approaches and practical guidelines* (3rd ed.). Boston: Pearson.

Grouws, D., & Cebulla, K. (2000). *Improving student achievement in mathematics*. Brussels: International Academy of Education.

Hope, M. (2007). Mathematical literacy. *Principal Leadership, 795,* 28-31.

Kaser, J. S., & Bourexis, P. S. (with Loucks-Horsley, S., & Raisen, S. A.). (1999). *Enhancing program quality in science and mathematics*. Thousand Oaks, CA: Corwin.

Kleinbaum, S. G, Kupper, L. L., & Muller, K.E. (1988). *Applied regression analysis and other multivariable methods* (2nd ed.). Boston: PWS-KENT Publishing Company.

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

Rossi, P. H., Lipsey, M. W. & Freeman, H. E. (2003). *Evaluation: A systematic approach* (7th ed.). Thousand Oaks, CA: Sage Publications.

Schorr, R. Y., & Amit, M. (2005). Analyzing student modeling cycles in the context of a 'real world' problem. In Chick, H. L. & Vencent, J. L. (Eds.). *Proceedings of the 29th conference of the international group for the psychology of mathematics education* (Vol. 4, pp. 137-144). Melbourne, Australia: PME.

Sood, S., & Jitendra, A. K. (2007). A comparative analysis of number sense instruction in reform-based and traditional mathematics textbooks. *The Journal of Special Education, 41*(3), 145-156.

Appendix

Table A1

Baseline Math Scores—Full and Stratified By Demographics—Grade 3

Characteristics	Type of score						
	Total			Multiple-choice		Constructed-response	
	N	Mean	*SD*	Mean	*SD*	Mean	*SD*
Full sample	265	8.95	3.67	6.98	2.76	1.97	1.52
Ethnicity							
White	144	9.33	3.53	7.19	2.64	2.13	1.51
Black	50	7.16	3.17	5.68	2.44	1.48	1.50
Hispanic	23	7.78	3.04	6.39	2.21	1.39	1.47
Asian American	48	10.23	4.11	7.98	3.19	2.25	1.48
Gender							
Male	143	9.24	3.71	7.23	2.75	2.01	1.53
Female	122	8.59	3.62	6.69	2.76	1.91	1.52
SES							
Low	33	7.27	3.44	5.94	2.61	1.33	1.38
Other	232	9.19	3.65	7.13	2.76	2.06	1.52
LEP							
Yes	9	7.67	3.64	5.22	1.92	2.44	2.07
No	256	8.99	3.68	7.04	2.77	1.95	1.50

Note. LEP = limited English proficient (term used in the Lawrence school district database), SES = socio-economic status.

Table A2

Baseline Math Scores—Full and Stratified By Demographics—Grade 4

Characteristics	Type of score						
	Total			Multiple-choice		Constructed-response	
	N	Mean	*SD*	Mean	*SD*	Mean	*SD*
Full sample	242	12.68	4.47	9.06	3.01	3.62	2.06
Ethnicity							
White	134	13.53	4.40	9.45	2.99	4.08	2.02
Black	44	10.32	3.90	7.91	2.86	2.41	1.86
Hispanic	27	9.78	3.81	7.37	2.68	2.41	1.67
Asian American	37	14.54	3.78	10.27	2.60	4.27	1.77
Gender							
Male	119	12.34	4.46	8.93	2.90	3.41	2.07
Female	123	13.01	4.48	9.19	3.12	3.82	2.05
SES							
Low	41	10.29	4.25	7.76	3.07	2.54	1.90
Other	201	13.17	4.37	9.33	2.93	3.84	2.03
LEP							
Yes	2	7.50	3.54	5.00	2.83	2.50	0.71
No	240	12.73	4.46	9.10	2.99	3.63	2.07

Note. LEP = limited English proficient (term used in the Lawrence school district database), SES = socio-economic status.

Table A3

Baseline Math Scores—Full and Stratified By Demographics—Grade 5

Characteristics	Type of score						
		Total		Multiple-choice		Constructed-response	
	N	Mean	*SD*	Mean	*SD*	Mean	*SD*
Full sample	260	11.27	4.94	7.87	3.15	3.40	2.25
Ethnicity							
White	143	11.81	4.56	8.33	2.80	3.48	2.22
Black	56	9.21	4.40	6.50	2.92	2.71	1.90
Hispanic	28	9.68	4.19	6.50	2.65	3.18	2.16
Asian American	33	13.73	6.33	9.36	4.11	4.36	2.68
Gender							
Male	147	11.54	5.02	8.09	3.18	3.46	2.32
Female	113	10.90	4.83	7.58	3.11	3.32	2.17
SES							
Low	44	9.89	5.04	6.82	3.15	3.07	2.32
Other	216	11.55	4.89	8.08	3.12	3.46	2.24
LEP							
Yes	10	7.20	4.34	5.00	3.02	2.20	1.48
No	250	11.43	4.90	7.98	3.11	3.44	2.27

Note. LEP = limited English proficient (term used in the Lawrence school district database), SES = socio-economic status.

Table A4

NJ ASK—Full and Stratified by Ethnicity—Grade 3

	Total	White	Black	Hispanic	Asian American	Possible
			Raw total			
Mean	22.70	23.87	18.55	19.84	25.06	33
SD	5.86	5.40	5.43	5.59	5.14	
			Scaled			
Mean	238.76	243.48	222.12	227.74	247.91	300
SD	23.28	20.64	25.04	23.30	19.39	
			Numbers			
Mean	6.42	6.77	5.17	5.72	7.06	9
SD	1.79	1.70	1.71	1.78	1.42	
			Geometry & measurement			
Mean	5.55	8.76	4.66	4.74	6.29	8
SD	1.93	1.54	1.39	1.32	1.47	
			Patterns & algebra			
Mean	5.64	5.88	4.78	4.96	6.18	8
SD	1.84	1.81	1.74	1.80	1.70	
			Data analysis			
Mean	5.10	5.46	3.94	4.48	5.53	8
SD	1.93	1.89	1.79	1.78	1.74	
			Problem solving			
Mean	7.52	8.04	5.42	6.52	8.71	12
SD	2.98	2.73	2.86	2.78	2.73	
N	265	143	50	23	45	

Table A5

NJ ASK—Full and Stratified by Ethnicity—Grade 4

	Total	White	Black	Hispanic	Asian American	Possible
Raw total						
Mean	25.41	27.17	19.29	18.33	31.30	43
SD	8.94	8.19	8.42	7.46	5.73	
Scaled						
Mean	231.32	237.87	208.91	205.74	252.32	300
SD	32.17	28.57	32.48	29.73	17.71	
Numbers						
Mean	8.93	9.44	7.45	6.93	10.30	13
SD	2.76	2.43	2.98	2.73	2.14	
Geometry & measurement						
Mean	5.25	5.54	3.74	3.96	6.92	10
SD	2.38	2.29	2.06	1.81	1.98	
Patterns & algebra						
Mean	6.27	6.63	4.93	4.37	7.86	10
SD	2.68	2.53	2.59	2.71	1.75	
Data analysis						
Mean	4.95	5.56	3.16	3.07	6.22	10
SD	2.65	2.52	2.45	2.00	1.95	
Problem solving						
Mean	13.49	14.63	9.67	9.74	16.54	23
SD	5.56	5.09	5.56	4.96	3.76	
N	241	134	43	27	37	

Table A6

NJ ASK—Full and Stratified by Ethnicity—Grade 5

	Total	White	Black	Hispanic	Asian American	Possible
			Raw total			
Mean	25.01	26.27	20.77	22.63	28.76	39
SD	6.92	6.27	6.85	6.56	6.16	
			Scaled			
Mean	231.70	237.77	211.27	219.33	250.36	300
SD	34.23	31.19	33.44	32.28	31.20	
			Numbers			
Mean	7.04	7.54	5.61	6.44	7.79	10
SD	2.26	2.06	2.15	2.55	1.87	
			Geometry & measurement			
Mean	6.17	6.32	5.52	5.96	6.82	9
SD	1.66	1.62	1.72	1.60	1.47	
			Patterns & algebra			
Mean	6.63	6.85	5.55	5.96	8.09	10
SD	2.15	1.99	2.18	2.21	1.67	
			Data analysis			
Mean	5.17	5.56	4.09	4.26	6.06	10
SD	2.31	2.23	2.04	1.89	2.57	
			Problem solving			
Mean	15.04	16.01	11.93	13.41	17.48	25
SD	4.80	4.40	4.65	4.31	4.18	
N	258	142	56	27	33	

Table A7

Adjusted and Unadjusted Effects of Math Out of the Box (MTB) on Math Achievement: ETS Assessment

Type	Adjusted means				Unadjusted means			
	MTB	Non-MTB	Diff	*p*-value	MTB	Non-MTB	Diff	*p*-value
			Grade 3					
Total	17.53	16.13	1.40	.001	17.57	16.10	1.47	.006
Multiple-choice	12.51	11.72	0.79	.010	12.58	11.67	0.91	.009
Constructed-response	4.98	4.44	0.55	.016	4.99	4.99	0.56	.029
			Grade 4					
Full	15.04	13.12	1.92	.000	16.66	12.30	4.36	.000
Multiple-choice	11.83	9.98	1.85	.000	12.80	9.49	3.31	.000
Constructed-response	3.46	3.02	0.44	.074	3.86	2.81	1.05	.000
			Grade 5					
Total	17.59	14.57	3.02	.000	16.72	14.71	2.01	.060
Multiple-choice	12.35	10.46	1.88	.000	11.80	10.55	1.25	.067
Constructed-response	5.21	4.11	1.10	.006	4.92	4.16	0.76	.122

Table A8

Adjusted and Unadjusted Effects of Math Out of the Box (MTB) on Math Achievement: NJ ASK

Type	Adjusted means				Unadjusted means			
	MTB	Non-MTB	Diff	*p*-value	Non-MTB	MTB	Diff	*p*-value
Grade 3								
Total (raw)	23.35	22.27	1.08	.04	23.41	22.23	1.18	.11
Number operations	6.49	6.37	0.12	.47	6.52	6.35	0.17	.45
Math patterns	5.86	5.48	0.38	.05	5.84	5.50	0.35	.13
Problem solving	7.85	7.29	0.56	.06	7.88	7.27	0.37	.10
Grade 4								
Total (raw)	26.42	24.90	1.52	.07	30.42	22.90	7.52	.0001
Number operations	9.25	8.78	0.47	.14	10.26	8.27	1.99	.0001
Math patterns	6.40	6.20	0.20	.48	7.54	5.63	1.92	.0001
Problem solving	14.09	13.19	0.90	.09	16.54	11.97	4.58	.0001
Grade 5								
Total (raw)	25.82	24.89	.93	.24	25.04	24.80	0.25	.86
Number operations	7.10	7.03	.07	.83	7.08	6.80	2.8	.50
Math patterns	6.85	6.60	.25	.42	6.63	6.66	-0.03	.94
Problem solving	15.10	15.03	.07	.90	15.14	14.40	0.74	.39