metis associates
making a meaningful difference

120 Wall Street
21st Floor
New York, NY 10005
Tel: 212-425-8833
Fax: 212-480-2176
www.metisassociates.com

Classroom, Inc.
Combined 2008-2011 MLI and RLI Summer Test Results
Prepared by Metis Associates
November 2011

To assess student performance in reading and mathematics, teachers implementing the Classroom, Inc. (CI) program during the past four summers (2008-2011) administered a reading assessment – the Reading-Level Indicator (RLI) – and/or a math assessment - the Math-Level Indicator (MLI) – before and after the program was used (pretest/posttest design). These assessments were developed by the American Guidance Service, Inc. (AGS) and have been documented to be valid and reliable measures of students' reading and math skills, when used for program evaluation.

For the past four years, from 2008 through 2011, Classroom, Inc. has contracted with Metis Associates to analyze and report on the reading and mathematics performance of students participating in CI's summer programming. This report, in turn, combines the data from these four summers into a comprehensive dataset for students who took the RLI (N=2,398) and students who took the MLI (N=1,818).

Using this new dataset, paired-samples *t*-tests were conducted on the scale scores[1] to determine whether there were statistically significant differences in students' reading and mathematics scores from pretest to posttest. Results were disaggregated by funder and/or partner, grade level and gender. Effect sizes were also calculated using Cohen's *d* to determine the magnitude of the differences. According to Cohen (1988), effect sizes of .2 are generally small, .5 are medium and .8 are large. In addition, to determine whether outcomes were associated with the intensity of program use (dosage), analyses of covariance were conducted to examine whether the number of episodes students completed was associated with more positive student outcomes as indicated by higher test scores. Finally, changes in grade level equivalents based on the mean pretest and posttest scale scores in reading and mathematics are presented by grade level.

RLI Results

From 2008 to 2011, a total of 2,434 students[2] completed both a pretest and posttest reading assessment. Thirty-six students were excluded from the analyses because they had completed less than half of the items on either the pretest or posttest. Overall, participating students showed no summer learning loss, which for students from low-income areas is typically a two-month loss over the summer months. Instead, students using the CI program for 4 to 5 weeks experienced gains in their reading performance from an average pretest scale score of 112.29 to an average posttest scale score of 113.35. Gains were statistically significant at the .05 level. This finding, along with the overall changes in reading

[1] To prepare for these pre/post analyses, raw test scores were converted into scale scores following the W-ability scale score conversion provided in the American Guidance Service, Inc. test manuals. The scale scores of students who responded to less than half of the items in either the pretest or the posttest were excluded from these analyses.
[2] Note that no longitudinal analyses were conducted: students who may have participated in more than one year are treated as separate individuals for these analyses.

performance for each year (2008-2011), is presented in Table 1. Note that from 2009 through 2011, students exhibited statistically significant gains in reading performance over the summer.

Table 1
Summers 2008-2011 Combined Analyses
Reading-Level Indicator Test Results
Changes in Scale Scores, Across Years and by Implementation Year

Implementation Year	Matched N	Pretest Scale Score	Posttest Scale Score	Change in Scores (Posttest-Pretest)	*t* (Sig.) [a]	Effect Size (*Cohen's d*) [b]
2008	160	112.43	112.88	0.45	0.784 (.434)	-
2009	942	110.40	111.28	0.88	3.465 (.001)*	0.11
2010	896	111.80	113.09	1.29	5.575 (.000)*	0.19
2011	400	117.82	118.98	1.16	3.361 (.001)*	0.17
Total (2008-2011)	2398	112.29	113.35	1.05	7.051 (.000)*	0.14

[a] An asterisk in this column denotes a statistically significant difference at the p≤.05 level based on a paired-samples *t*-test.
[b] Effect size is a measure of the magnitude of the gains or losses, expressed in gain score standard deviation units.

The data were disaggregated by the geographic location of Classroom, Inc.'s partners/funders. These results reveal that students in New York City, Chicago, Newark, and Memphis schools experienced statistically significant gains in average scale scores from pretest to posttest. Table 2 presents these results.

Table 2
Summers 2008-2011 Combined Analyses
Reading-Level Indicator Test Results
Changes in Scale Scores by Geographic Location of Funder/Partner sites

Location	Matched N	Pretest Scale Score	Posttest Scale Score	Change in Scores (Posttest-Pretest)	*t* (Sig.) [a]	Effect Size (*Cohen's d*) [b]
New York City	606	116.51	117.46	0.94	3.103 (.002)*	0.13
Chicago	1104	110.60	111.77	1.17	5.44 (.000)*	0.16
Newark	587	110.62	111.56	0.94	2.983 (.003)*	0.12
Memphis	101	115.29	116.31	1.02	1.856 (.066)	-

[a] An asterisk in this column denotes a statistically significant difference at the p≤.05 level based on a paired-samples *t*-test.
[b] Effect size is a measure of the magnitude of the gains or losses, expressed in gain score standard deviation units.

The data also were disaggregated by grade level and gender. As shown in Table 3, a statistically significant gain was observed among students in grades 5 through 9, who accounted for 93 percent of all students. These gains represented small to moderate changes. When looking at results by gender, the results also show that male and female students each experienced statistically significant gains.

Table 3
Summers 2008-2011 Combined Analyses
Reading-Level Indicator Test Results
Changes in Scale Scores by Grade Level and Gender

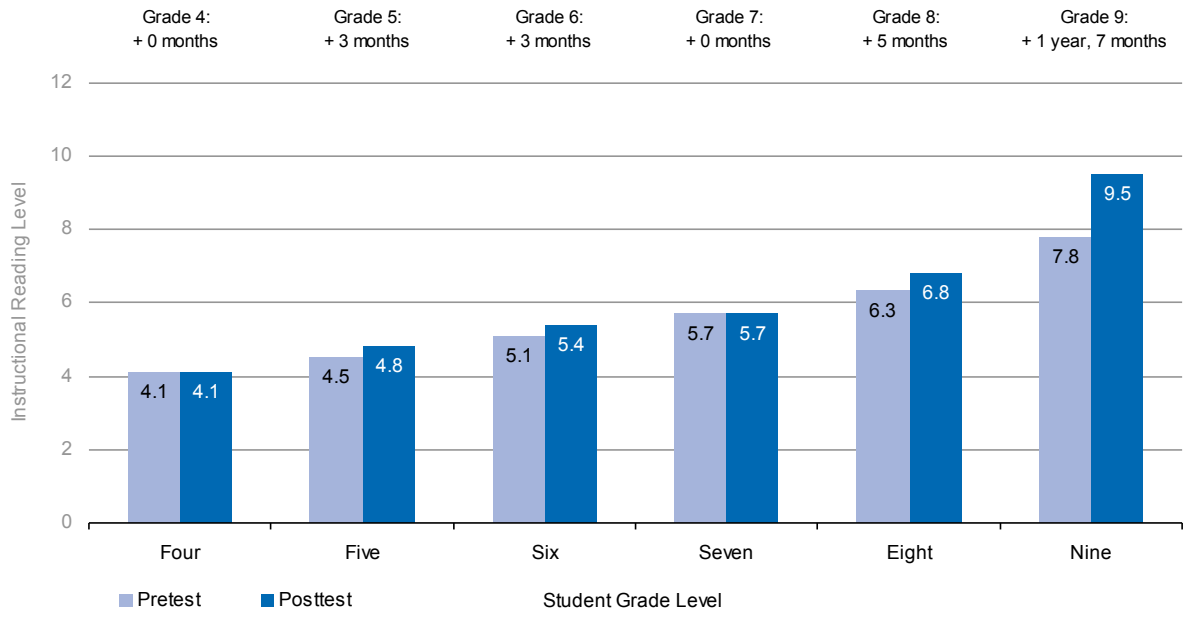
Student Characteristics		Matched N[a]	Pretest Scale Score	Posttest Scale Score	Change in Scores (Posttest-Pretest)	*t* (Sig.)[b]	Effect Size (*Cohen's d*) [c]
Grade	Three	3	-	-	-	-	-
	Four	166	104.15	104.69	0.55	0.864 (.389)	-
	Five	194	108.03	109.05	1.02	2.121 (.035)*	0.15
	Six	724	110.48	111.96	1.48	5.689 (.000)*	0.21
	Seven	442	112.98	113.82	0.84	2.416 (.016)*	0.11
	Eight	581	114.87	115.79	0.92	2.829 (.005)*	0.12
	Nine	252	118.86	119.83	0.97	2.315 (.021)*	0.15
Gender	Male	1365	112.44	113.47	1.02	5.178 (.000)*	0.14
	Female	999	112.14	113.31	1.17	5.074 (.000)*	0.16

[a] Statistical analyses were not conducted for students in grade three due to the small sample size.
[b] An asterisk in this column denotes a statistically significant difference at the $p \leq .05$ level based on a paired-samples *t*-test.
[c] Effect size is a measure of the magnitude of the gains or losses, expressed in gain score standard deviation units.

To help in understanding the results, the average scale scores at pretest and posttest were converted into grade equivalents (GE) of students' instructional reading level.[3] As shown in Figure 1, the results show higher instructional reading levels for the posttest mean scores than for the pretest mean scores for grades 5, 6, 8, and 9. The results show an approximate gain of three months in grades 5 and 6, and larger gains were observed for grades 8 and 9 (five months and one year and seven months accordingly). For example, the average reading scale score for eighth-grade students increased from 114.87 at pretest, which represents a sixth grade, third month reading level, to 115.79 at posttest, which represents a sixth grade, eighth month reading level. The average scale score for ninth-grade students increased from 118.86 in the pretest, which corresponds to a seventh grade, eighth month reading level, to 119.83 in the posttest, which corresponds to a ninth grade, fifth month reading level. Table A2 in the appendix presents the results of the grade-equivalent analyses by funder/partner as well.

[3] As described in the publisher's manual, "grade equivalents are referred to as *developmental norms* because they place an individual along a span or continuum of development. Grade-equivalent values are presented in tenths of a grade." To calculate GEs, average scale scores were first converted to raw scores and then into a GE following the publisher's conversion tables provided in the manual. The average scale score across all students (112.29) at pretest corresponds to a raw score of 23 which corresponds to a GE value of 5.4; the average scale score (113.35) at posttest corresponds to a raw score of 24 which corresponds to a GE value of 5.7. Note that GEs have many limitations. Since they are not equal-interval scales of measurement, they cannot be manipulated arithmetically (e.g., averaged) or used for direct longitudinal comparisons.

Figure 1
Summers 2008-2011 Combined Analyses
Reading-Level Indicator Test Results
Changes in Grade Level Equivalencies by Grade Level



Students and teachers were asked to report the number of episodes that they and their pupils completed. Of the 2,398 students with complete matched reading test data, data on the number of episodes were available for 1,975 students, representing 82 percent of the entire population.

Analyses of covariance were conducted to determine whether there were differences in posttest scores across groups of students who completed 0 to 4 episodes, 5 to 7 episodes, or 8 or more episodes, after controlling for pretest scores. As seen in Table 4, although students who completed 8 or more episodes had higher adjusted mean test scores (after taking into account pretest differences) than students in the two lower dosage groups, these differences were not statistically significant at the .05 level. The correlation between the number of episodes completed and students' pre/post gain scores was not found to be statistically significant.

Table 4
Summers 2008-2011 Combined Analyses
Reading-Level Indicator Test Results
Reading-Level Indicator Test Results, by Degree of Program Implementation

Number of Episodes	Total N	Posttest Adjusted Mean Score[1]	F (Sig.)[2]	Effect size[3]	Post Hoc Comparisons
0 to 4	413	113.87	2.435 (.088)	-	-
5 to 7	1045	113.35			
8 or more	517	114.11			

[1] Posttest mean scores were adjusted to take into account pretest differences in W-ability scores.
[2] An asterisk denotes a statistically significant difference at the .05 level based on an analysis of covariance.
[3] Effect size is a measure of the magnitude of the gains or losses, expressed in gain score standard deviation units.

MLI Results

From 2008 to 2011, a total of 1,839 students[4] completed both a pretest and posttest mathematics assessment. Twenty-one students were excluded from the analyses because they had completed less than half the items on either the pretest or the posttest. Overall, the results revealed no summer learning loss in math, which should be considered a very positive finding since research has typically shown performance losses during the summer months. Specifically, when assessing changes across all four summers combined, the average scale score increased from 104.59 in the pretest to 107.11 in the posttest. These gains were statistically significant beyond the .01 level. This finding, along with the overall changes in mathematics performance for each year (2008-2011), is presented in Table 5. The results also show that for three out of the last four years (from 2008 through 2010), students have exhibited statistically significant gains in mathematics performance over the 4 to 5 week Classroom, Inc. program.

Table 5
Summers 2008-2011 Combined Analyses
Mathematics-Level Indicator Test Results
Changes in Scale Scores, Across Years and by Implementation Year

Implementation Year	Matched N	Pretest Scale Score	Posttest Scale Score	Change in Scores (Posttest-Pretest)	*t* (Sig.) [a]	Effect Size (*Cohen's d*) [b]
2008	91	107.68	109.74	2.06	3.089 (.003)*	0.32
2009	738	103.25	106.00	2.75	11.556 (.000)*	0.43
2010	852	104.01	106.72	2.71	12.164 (.000)*	0.42
2011	137	113.39	113.77	0.38	0.643 (.521)	-
Total (2008-2011)	1818	104.59	107.11	2.51	16.423 (.000)*	0.39

[a] An asterisk in this column denotes a statistically significant difference at the p≤.05 level based on a paired-samples *t*-test.
[b] Effect size is a measure of the magnitude of the gains or losses, expressed in gain score standard deviation units.

The data also were analyzed by the geographic location of Classroom, Inc's partner/funders. Table 6 presents these results. When looking at the findings by location, the results reveal that Chicago and Newark students experienced statistically significant gains in average scale scores from pretest to posttest. New York City students also experienced gains in their mathematics scores, although the gains did not reach statistical significance.

Table 6
Summers 2008-2011 Combined Analyses
Mathematics-Level Indicator Test Results
Changes in Scale Scores by Funder/Partner

Location	Matched N	Pretest Scale Score	Posttest Scale Score	Change in Scores (Posttest-Pretest)	*t* (Sig.)[a]	Effect Size (*Cohen's d*)[b]
New York City	229	111.34	112.11	0.78	1.784 (.076)	-
Chicago	1061	103.23	106.06	2.83	14.082 (.000)*	0.43
Newark	528	104.40	107.04	2.64	9.518 (.000)*	0.41

[a] An asterisk in this column denotes a statistically significant difference at the p≤.05 level based on a paired-samples *t*-test.
[b] Effect size is a measure of the magnitude of the gains or losses, expressed in gain score standard deviation units.

[4] Note that no longitudinal analyses were conducted: students who may have participated in more than one year are treated as separate individuals for these analyses.

Results also were disaggregated by grade level and gender. These results, presented in Table 7, indicate that students in the fourth through eighth grades experienced statistically significant gains in mathematics performance. Students in these grades account for 93 percent of all students (with available grade information). When looking at gender, results also show that male and female students each experienced statistically significant gains in their math performance.

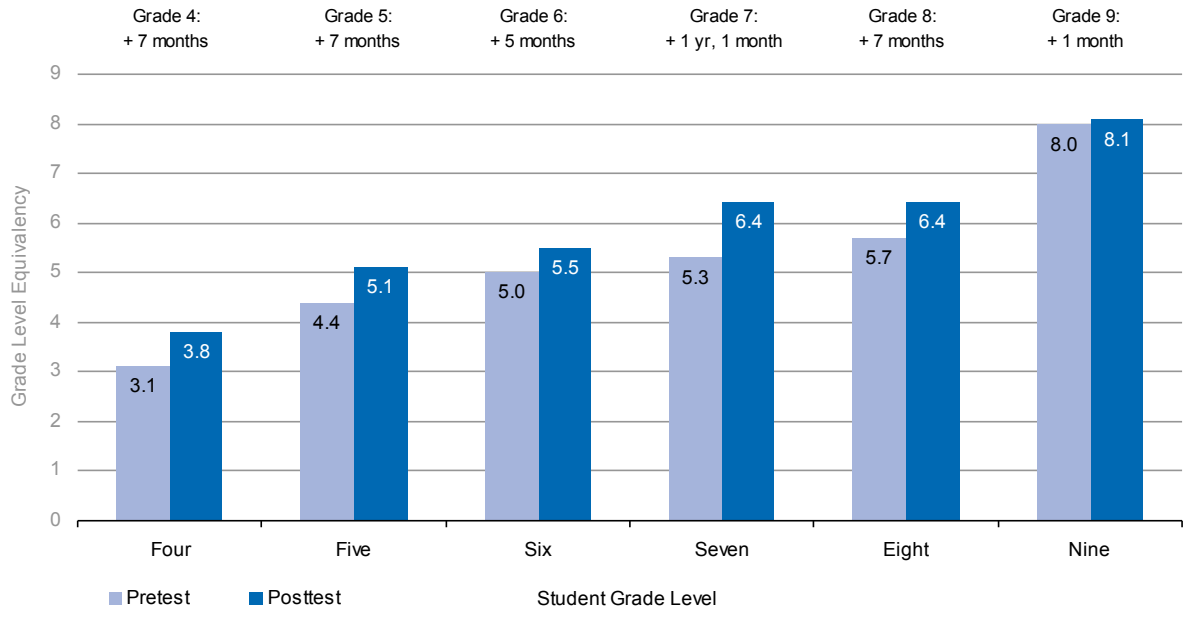
Table 7
Summers 2008-2011 Combined Analyses
Mathematics-Level Indicator Test Results
Changes in Scale Scores by Grade Level and Gender

Student Characteristics		Matched N[a]	Pretest Scale Score	Posttest Scale Score	Change in Scores (Posttest-Pretest)	*t* (Sig.)[b]	Effect Size (*Cohen's d*)[c]
Grade	Three	7	-	-	-	-	-
	Four	129	97.43	101.47	4.04	6.722 (.000)*	0.59
	Five	152	102.08	105.14	3.06	5.926 (.000)*	0.48
	Six	546	103.50	106.34	2.84	10.824 (.000)*	0.46
	Seven	334	105.26	107.60	2.34	6.02 (.000)*	0.33
	Eight	498	106.24	108.42	2.17	7.332 (.000)*	0.33
	Nine	122	112.73	113.41	0.68	1.342 (.182)	-
Gender	Male	1042	104.34	106.88	2.53	12.387 (.000)*	0.38
	Female	750	104.97	107.50	2.52	10.76 (.000)*	0.39

[a] Statistical analyses were not conducted for students in grade three due to the small sample size.
[b] An asterisk in this column denotes a statistically significant difference at the $p \leq .05$ level based on a paired-samples *t*-test.
[c] Effect size is a measure of the magnitude of the gains or losses, expressed in gain score standard deviation units.

To help in understanding the results, the average scale scores at pretest and posttest were converted into grade equivalents (GE). Figure 2 presents these data disaggregated by grade level. The results show an approximate gain of seven months in grades four, five, six, and eight, with a substantially larger gain of one year, one month in grade seven and a comparatively low gain of only one month in grade nine. The results are presented by funder/partner and school in Table A6 in the Appendix.

Figure 2
Summers 2008-2011 Combined Analyses
Mathematics-Level Indicator Test Results
Changes in Grade Level Equivalencies by Grade Level



Students and teachers were asked to report the number of simulation episodes they and their pupils completed. Of the 1,818 students with complete matched test data, data on the number of episodes completed was available for 1,491 students, representing 82 percent of the population.

Analyses of covariance were conducted to determine whether there were differences in posttest scores across groups of students that completed 0 to 4 episodes, 5 to 7 episodes, or 8 or more episodes, after controlling for pretest scores. As seen in Table 8, there were no statistically significant differences in posttest test adjusted mean scores across these three groups. However, a positive correlation[5] was found between the number of episodes completed and the change in pre/post mathematics performance.

Table 8
Summers 2008-2011 Combined Analyses
Mathematics-Level Indicator Test Results
Math-Level Indicator Test Results, by Degree of Program Implementation

Number of Episodes	Total N	Posttest Adjusted Mean Score[1]	F (Sig.)[2]	Effect size[3]	Post Hoc Comparisons
0 to 4	227	106.73	1.796 (.166)	-	-
5 to 7	717	107.57			
8 or more	547	107.54			

[1] Posttest mean scores were adjusted to take into account pretest differences in W-ability scores.
[2] An asterisk denotes a statistically significant difference at the .05 level based on an analysis of covariance.
[3] Effect size is a measure of the magnitude of the gains or losses. Effect sizes of .2 are considered small, .5 are considered medium, and .8 are considered large.

[5] *rho* = .07, p = .007

Conclusions

The results from the combined summer 2008-2011 analyses indicate that there were statistically significant gains in the reading and mathematics performance of students who participated in a 4 to 5 week summer program with Classroom, Inc. This is especially noteworthy as research in summer learning has typically shown that low income students exhibit a performance loss during the summer, often leading to a wider achievement gap between students of lower and higher socioeconomic backgrounds. The results also indicate that students improved their reading performance in all four cities (New York City, Newark, Memphis, and Chicago) where the CI program took place over the past four summers. Similarly, students improved their mathematics performance in two (Newark and Chicago) of the three cities where the CI program took place over the past four summers. And, in New York City, students also experienced gains in their mathematics scores, although the gains did not reach statistical significance. In addition, a statistically significant correlation was found between the number of episodes students completed and their gain scores in mathematics. Overall, improvements in math performance were found to be stronger than those in reading, both in effect size (.39 compared to .14) and change in average grade level equivalency among students (a gain of seven months for math compared to a gain of three months for reading).

Appendix

Table A1
Summers 2008-2011 Combined Analyses
Reading-Level Indicator Test Results
Instructional Reading Level (IRL) by Funder/Partner

Implementation Year	Matched N	Pretest Scale Score	Posttest Scale Score	Pretest IRL	Posttest IRL	IRL Change
2008	160	112.43	112.88	5.4	5.7	+3 months
2009	942	110.40	111.28	5.1	5.4	+3 months
2010	896	111.80	113.09	5.4	5.7	+3 months
2011	400	117.82	118.98	7.8	7.8	+0 months
Total (2008-2011)	2398	112.29	113.35	5.4	5.7	+3 months

Table A2
Summers 2008-2011 Combined Analyses
Reading-Level Indicator Test Results
Instructional Reading Level (IRL) by Implementation year

Location	Matched N	Pretest Scale Score	Posttest Scale Score	Pretest IRL	Posttest IRL	IRL Change
New York City	606	116.51	117.46	6.9	7.8	+11 months
Chicago	1104	110.60	111.77	5.1	5.4	+3 months
Newark	587	110.62	111.56	5.1	5.4	+3 months
Memphis	101	115.29	116.31	6.3	6.8	+5 months

Table A3
Summers 2008-2011 Combined Analyses
Reading-Level Indicator Test Results
Instructional Reading Level (IRL) by Student Characteristics

Student Characteristics		Matched N	Pretest Scale Score	Posttest Scale Score	Pretest IRL	Posttest IRL	IRL Change
Grade	Three	3	-	-	-	-	-
	Four	166	104.15	104.69	4.1	4.1	+0 months
	Five	194	108.03	109.05	4.5	4.8	+3 months
	Six	724	110.48	111.96	5.1	5.4	+3 months
	Seven	442	112.98	113.82	5.7	5.7	+0 months
	Eight	581	114.87	115.79	6.3	6.8	+5 months
	Nine	252	118.86	119.83	7.8	9.5	+1 year, 7 months
Gender	Male	1365	112.44	113.47	5.4	5.7	+3 months
	Female	999	112.14	113.31	5.4	5.7	+3 months

Table A4
Summers 2008-2011 Combined Analyses
Reading-Level Indicator Test Results
Changes in Scale Scores from Pretest to Posttest

Population		Matched N[a]	Changes from Pretest to Posttest		
			% Declined	% No Change	% Improved
Location	New York City	606	40.3%	5.3%	54.5%
	Chicago	1104	44.5%	3.6%	51.9%
	Newark	587	45.0%	3.7%	51.3%
	Memphis	101	42.6%	5.9%	51.5%
Implementation Year	2008	160	44.4%	6.3%	49.4%
	2009	942	44.2%	4.0%	51.8%
	2010	896	43.4%	3.6%	53.0%
	2011	400	41.5%	5.0%	53.5%
Grade	3	3	-	-	-
	4	166	51.2%	2.4%	46.4%
	5	194	45.9%	2.6%	51.5%
	6	724	43.1%	3.7%	53.2%
	7	442	45.9%	4.5%	49.5%
	8	581	38.9%	5.5%	55.6%
	9	252	42.9%	4.4%	52.8%
Gender	Male	1365	43.8%	3.6%	52.6%
	Female	999	42.5%	4.9%	52.6%
All Participating Students 2008-2011[b]		2398	43.5%	4.21%	52.4%

[a] Statistical analyses were not conducted for students in grade three due to the small sample size.
[b] The matched N for this population exceeds the totals of matched students presented for location, implementation year, grade, and gender, because of missing data for those individual characteristics. Differences in the total Matched N for each category are within 40 students—less than 2% of the 2398 students presented in this grand total.

Table A5
Summers 2008-2011 Combined Analyses
Mathematics-Level Indicator Test Results
Mathematics Grade Level Equivalencies by Implementation year

Implementation Year	Matched N	Pretest Scale Score	Posttest Scale Score	Pretest GE	Posttest GE	GE Change
2008	91	107.68	109.74	6.4	7.2	+ 8 months
2009	738	103.25	106.00	4.7	5.5	+ 8 months
2010	852	104.01	106.72	5.0	6.0	+ 1 year
2011	137	113.39	113.77	8.0	8.5	+ 5 months
Total (2008-2011)	1818	104.59	107.11	5.3	6.0	+ 7 months

Table A6
Summers 2008-2011 Combined Analyses
Mathematics-Level Indicator Test Results
Mathematics Grade Level Equivalencies by Funder/Partner

Location	Matched N	Pretest Scale Score	Posttest Scale Score	Pretest GE	Posttest GE	GE Change
New York City	229	111.34	112.11	7.2	7.8	+6 months
Chicago	1061	103.23	106.06	4.7	5.5	+8 months
Newark	528	104.40	107.04	5.0	6.0	+1 year

Table A7
Summers 2008-2011 Combined Analyses
Mathematics-Level Indicator Test Results
Mathematics Grade Level Equivalencies by Student Characteristics

Student Characteristics		Matched N	Pretest Scale Score	Posttest Scale Score	Pretest GE	Posttest GE	GE Change
Grade	Three	7	-	-	-	-	-
	Four	129	97.43	101.47	3.1	3.8	+ 7 months
	Five	152	102.08	105.14	4.4	5.1	+ 7 months
	Six	546	103.50	106.34	5.0	5.5	+ 5 months
	Seven	334	105.26	107.60	5.3	6.4	+ 1 year, 1 month
	Eight	498	106.24	108.42	5.7	6.4	+ 7 months
	Nine	122	112.73	113.41	8.0	8.1	+ 1 month
Gender	Male	1042	104.34	106.88	5.0	6.0	+ 1 year
	Female	750	104.97	107.50	5.3	6.4	+ 1 year, 1 month

Table A8
Summers 2008-2011 Combined Analyses
Math-Level Indicator Test Results
Changes in Scale Scores from Pretest to Posttest

Population		Matched N[a]	Changes from Pretest to Posttest		
			% Declined	% No Change	% Improved
Location	New York City	229	38.9%	9.6%	51.5%
	Chicago	1061	28.9%	6.8%	64.3%
	Newark	528	30.1%	5.9%	64.0%
Implementation Year	2008	91	29.7%	8.8%	61.5%
	2009	738	29.5%	6.1%	64.4%
	2010	852	29.7%	7.0%	63.3%
	2011	137	41.6%	8.8%	49.6%
Grade	3	7	28.6%	0.0%	71.4%
	4	129	22.5%	7.0%	70.5%
	5	152	25.7%	7.2%	67.1%
	6	546	29.7%	5.9%	64.5%
	7	334	31.4%	7.2%	61.4%
	8	498	31.7%	6.4%	61.8%
	9	122	40.2%	11.5%	48.4%
Gender	Male	1042	29.8%	7.7%	62.5%
	Female	750	31.3%	5.7%	62.9%
All Participating Students 2008-2011[b]		1818	30.5%	6.9%	62.6%

[a] Statistical analyses were not conducted for students in grade three due to the small sample size.

[b] The matched N for this population exceeds the totals of matched students presented for location, implementation year, grade, and gender, because of missing data for those individual characteristics. Differences in the total Matched N for each category are within 40 students—less than 2% of the 1818 students presented in this grand total.