



The Massachusetts Even Start Family Literacy Program
Analysis of Student MCAS Outcomes

October 2009



Contents

Executive Summary2

Introduction3

Methodology5

 Sample 5

 Analysis of Data 9

Findings12

 MCAS Performance Levels for Even Start and Comparison Group.....12

 Analysis of MCAS Raw Scores and Standard Raw Scores 13

 Analysis of MCAS Outcomes of Students from Currently Operating Programs 15

 MCAS Results by Age at Program Enrollment and Total Months Attended.....17

Appendix A: Comparison to Statewide Low-Income and LEP Students19

Appendix B: Additional MCAS Grade Levels22

Executive Summary

This report summarizes the findings of an analysis conducted by the University of Massachusetts Donahue Institute to estimate the impact of participation in the Even Start Family Literacy Program on students' subsequent academic performance on Massachusetts Comprehensive Assessment System (MCAS) tests. Building on an earlier MCAS study commissioned by ESE in 2006, this study used a non-experimental design to compare the performance of Even Start students on the grade 3 ELA and grade 4 ELA and Mathematics MCAS tests to that of a demographically matched comparison group.

Student-level data from three ESE data collection systems—the Student Information Management System (SIMS), MCAS, and the System for Managing Accountability and Results through Technology (SMARTT), which houses information on students participating in the Massachusetts Even Start program—were used to identify Even Start students and match each to a similar non-Even Start student *at the time of the student's enrollment in the Even Start program or the student's first appearance in the SIMS file if not enrolled in a Massachusetts public school at the time of Even Start participation*, on the basis of important demographic characteristics. In total, the sample included 607 Even Start students and 607 comparison group students.

Because the sample was pooled and students took their MCAS exams in different years, student raw scores were standardized by test and year to have a mean of zero and a standard deviation of one, an adjustment that is consistent with ESE's *Researcher's Guide to Massachusetts State Education Data*. These scores were compared to identify differences between Even Start and comparison group students. The scores of Even Start students were also compared to identify whether age at program entry and duration of time in the program were related to MCAS outcomes.

The key findings of the study were as follows:

- Children who participated in Even Start scored slightly better on their subsequent grade 3 and 4 MCAS exams than those in a comparison group, although differences in performance were small and not statistically significant.
- Students who attended a subset of four currently operating programs that survived a round of deep budget cuts in 2007 outperformed students in their comparison group. Although not statistically significant, the differences between Even Start and comparison students for this subset of programs considered well-implemented by ESE managers was far larger than for all programs as a whole, suggesting that implementation quality matters.
- Among Even Start students, age of entry into the program appeared to have a weak, negative relationship to subsequent MCAS performance, meaning that students who entered the program at a younger age tended to perform better than those who joined the program when they were older.

These findings are generally consistent with those of the 2006 study. Overall, the number of Even Start students able to be included in the present study greatly exceeded the number included in the number of students in the 2006 study, providing additional statistical power and confidence in the results.

Introduction

The Massachusetts Even Start Family Literacy Program provides comprehensive family literacy services to parents and their children. Designed to “help families overcome intergenerational cycles of undereducation and poverty,” Even Start targets families with low educational and socioeconomic backgrounds. The Massachusetts Department of Elementary and Secondary Education’s (ESE’s) Adult and Community Learning Services (ACLS) unit, which administers federal grants for the program, estimates that more than 4,000 of the state’s struggling families have participated in one of the state’s local Even Start programs since 1993.

This report summarizes the findings of an analysis conducted by the University of Massachusetts Donahue Institute using ESE-provided data to estimate the impact of Even Start program participation on students’ subsequent academic performance on the Massachusetts Comprehensive Assessment System (MCAS) tests.

The Comprehensive Family Literacy Model

The Massachusetts Family Literacy Consortium defines family literacy as “coordinated learning among different generations in the same family which helps both adults and children reach their full personal, social, and economic potential.” The Massachusetts Even Start Family Literacy Program is a comprehensive family literacy program, the model for which is one of integrated service delivery. Participating families receive academic and learning services in an environment that also provides additional family supports, such as child care, counseling, and other social services through community partnerships.

The five core academic learning services provided to all Massachusetts Even Start families are adult education, child education, parent education, interactive literacy, and regular home visits. Each program component, as outlined by ESE program standards, is summarized briefly below.

- **Adult Education:** Using curricula aligned with Massachusetts Curriculum Frameworks, participating parents receive adult basic education, which, depending on their educational needs, may comprise English for Speakers of Other Languages (ESOL) instruction and/or adult secondary education, such as pre-General Education Development (GED), GED, Adult Diploma Program (ADP), or External Diploma Program (EDP) training.
- **Child Education:** While parents participate in adult education, one or more participating children ages birth through seven receives developmentally appropriate education. Teachers are expected to use a variety of instructional strategies to enhance children’s development and work in partnership with families.
- **Parent Education:** Participating parents also engage in weekly educational opportunities where they learn ways they can support their children’s development, build effective parent–teacher relationships, and other aspects of effective parenting. According to the ESE, parent education is frequently integrated into the adult education classes, although time spent in parent education does not substitute for required adult education time.
- **Interactive Literacy:** At least once each week, parents engage in a guided literacy activity with their participating children. Activities are designed to encourage parents to engage in home literacy activities and more effectively support their children’s literacy skills development.
- **Home Visits:** In addition to classroom instruction, monthly home visits “link children’s learning in school to activities in the home and support learning within the context of [families’] daily lives.” Home visits are

intended to support parents in creating a literate home environment and frequently entail modeling of positive interactions, reinforcement, and the provision of supportive feedback by the home visitor.

Report Overview

The purpose of this study was to identify whether, controlling for child, family, and district characteristics commonly associated with academic performance, there is evidence that participation in the Even Start Family Literacy Program affects students' subsequent MCAS performance. The report that follows is organized into two sections. The first presents a detailed discussion of the research methods used in the study, including sample selection and analysis of data. The second presents the findings of the study and related discussion.

Methodology

To identify the effect of Even Start participation on the academic performance of students, a non-experimental study design was used to compare the performance of Even Start students on the grade 3 ELA and grade 4 ELA and Mathematics MCAS assessments with that of a matched cohort of students. The study used existing student-level data provided by ESE, including data from the Student Information Management System (SIMS), student MCAS performance data, and the System for Managing Accountability and Results through Technology (SMARTT) database, which houses information on students participating in the Massachusetts Even Start program.

Sample

To identify the potential impacts of the Even Start program on participating students, the MCAS results of Even Start students must be compared with those of a comparable group of students who did not participate in the program. Because the study design was non-experimental in nature and did not involve random assignment of students to treatment (Even Start) and comparison (non-Even Start) groups, a comparison group of students demographically similar to those in the Even Start program was constructed using ESE-provided SIMS data files.

Developing the sample, therefore, involved a two-stage process. First, Even Start students who could be matched to a valid grade 3 or grade 4 MCAS record had to be identified using the ESE-provided data sets. Second, a comparison group needed to be constructed by identifying a match for each eligible Even Start student. These processes are described in detail below. Ultimately, the sample comprised 1,214 students—607 Even Start participants and 607 comparison group students.

Even Start Students

Data from ESE's SMARTT data system was used to identify students who participated in a Massachusetts Even Start program between the 2001–2002 school year (SY02) and the 2007–2008 school year (SY08). Because the outcome measure of interest was a student's MCAS performance, this study focused on students who were old enough to have taken an MCAS exam, which is first administered to students in grade 3, as of spring 2008. As such, students born after September 1, 2000 were excluded from the study. In total, 1,035 of the 2,191 students appearing in the SMARTT files were born prior to this date.

To be included in the sample, students needed to be able to be matched to a state-assigned student identifier, or SASID number, in the SY02 through SY08 end-of-year SIMS files and a valid MCAS test record taken between spring 2002 and spring 2008. As Table 1 shows, 965 of the 1,035 students, or 93%, were able to be matched to a SASID through a combination of data merging and manual cross-checking of files.

Most of the students (687) were matched to a SASID by merging SIMS data files into the SMARTT database using a student's first name, last name, and date of birth as the relevant identification fields. Because nearly 34% of Even Start students born prior to September 1, 2000 were not matched to a SIMS record through this initial process, a second process was undertaken whereby researchers manually checked the SIMS files for each of the 348 Even Start students not matched through the initial merging process. Although time-consuming, this process allowed records to be identified in the event that formatting differences were present between the SMARTT and SIMS files, such as whether or not a hyphen was used for a compound name, or there were data entry errors in one of the files, such as the misspelling of a student's name or transposition of a date. In total, an additional 278 Even Start students were matched to a SASID through this secondary process.

Table 1: Even Start Students Born on or before September 1, 2000 Matched to a SASID

	Born on or before 9/1/2000
Total in SMARTT file	1,035
Matched student to SIMS through merging	687
Matched student to SIMS manually	278
Total Matched to a SIMS record	965
Unable to match student to a SASID	70

Of those matched to a SASID in SIMS, 628 were able to be linked to a subsequent grade 3 or grade 4 MCAS record. In cases where a student had taken an MCAS exam more than once, only his or her first score was considered. Students who had taken an MCAS alternate assessment or whose test record was incomplete due to an absence or other reason were excluded from the analysis, resulting in the removal of 21 of the 628 records. In total, 607 Even Start students were able to be matched to a valid grade 3 or grade 4 MCAS record and thus included in the sample of Even Start students.

Comparison Group

To identify a comparison group, each of the 607 Even Start students in the sample was matched to a similar non-Even Start student on the basis of important demographic characteristics *at the time of the student's enrollment in the Even Start program or, in the case of students not enrolled in a Massachusetts public school at the time of Even Start enrollment, the student's first appearance in the SIMS files.*¹ Matching at the time of enrollment, as opposed to at the time of MCAS testing, allowed researchers to identify a comparison student who had similar characteristics to the Even Start student had that student not participated in the Even Start program.² Students who appeared in the SIMS but who could not be matched to a valid grade 3 or grade 4 MCAS record were excluded from the sample of possible matches.

The matching process relied on a series of tiered match rules designed to find the best available match for the student on the basis of observed characteristics known to influence student achievement outcomes.³ If a student could not be matched using a match rule, a less restrictive match rule was used until a similar student could be found. To control for district characteristics, students were only matched to other students from the same town, unless a match could not be found there using any of the match rules, in which case a student was chosen from a comparable district of similar size, low-income and limited English proficiency (LEP) rates, and district performance on the grade 3 ELA MCAS exam (measured by ESE's composite performance index). Match rules were as follows:

¹ In order to ensure that Even Start students were not matched to other Even Start students, data from the ESE-provided SMARTT files was used to identify students in the SIMS who had participated in Even Start. These students were excluded from the universe of possible comparison group students.

² Program managers note that Even Start is intended to provide literacy training and other skills that lead to increased economic self-sufficiency for parents, which may cause some participating students and their families to move out of low-income status. Similarly, students who may be limited English proficient upon enrollment in Even Start may gain language and literacy skills, such that they no longer qualify to be considered an LEP student.

³ An alternate technique for matching, propensity score matching, was considered, but not deemed the best method for this particular study. Propensity score matching uses a multi-variable analysis to predict a probability, or propensity score, that a student would have received the treatment (i.e. participated in Even Start) on the basis of observed characteristics, and students are selected into the comparison group based on their propensity scores. This technique is most appropriate for large samples, and the number of Even Start students in each individual year was typically small.

Selection Rule 1: The student was matched to another student of the same race, gender, grade level, low-income profile, LEP classification, type of special education placement,⁴ and who spoke the same first language.

Selection Rule 2: The student was matched to another student using the criteria outlined in the previous rule, except a student whose first language was not English was able to be matched to another non-native English speaker whose first language may have been different from their own.

Selection Rule 3: The student was matched to another student using the criteria outlined in the previous rule, except special education students were matched on the basis of whether or not they participated in special education and not the type of program in which they were enrolled.

Selection Rule 4: The student was matched to another student using the criteria outlined in the previous rule, except gender was not considered.

Selection Rule 5: The student was matched to another student using the criteria outlined in the previous rule, except a non-white student could be matched to another student of a different race, provided that that student was also non-white.

Matches were selected without replacement, meaning that once a student was chosen as a match, that student was no longer eligible to be considered a match for other students. This prevented a single student from being considered as a match to multiple students.

Table 2 displays the number and proportion of Even Start students matched using each of the rules. Overall, the vast majority of students (89%) were able to be matched using the *Selection 1* match rules.

Table 2: Summary of Match Rules Used to Select Students

	N	%
Selection 1	540	89%
Selection 2	19	3%
Selection 3	6	1%
Selection 4	19	3%
Selection 5	11	2%
Matched to a similar student in a comparable district	12	2%
Total	607	100%

The demographic profile of Even Start students included in the sample (i.e., those students who were able to be matched to a grade 3 or grade 4 MCAS test) and those of students in the comparison group are shown in Table 3. As this table shows, the Even Start and the comparison groups were comparable in terms of all measured demographic characteristics, including race and gender, and special population classifications, including low-income, first language not English (FLNE), LEP, and special education statuses. It should be noted that many Even Start programs serve students too young to be enrolled in a Massachusetts public school at the time of Even Start participation (and consequent appearance in the SIMS data), and the program is designed to provide services that may cause a student who was LEP or low-income at the time of participation to move out of those categories. As such, it should be noted that LEP and low-income rates for the sample reflect the rates reported in SIMS at the time of the match and not rates at the time of enrollment in Even Start.

⁴ Specifically, for this level of a match, special education students were matched to students with a similar level of intervention—full inclusion, partial inclusion, substantially separate, out of district placement, or in the case of the SY02 and SY03 SIMS, preschool-aged students receiving special education services.

Table 3: Demographic Characteristics of Even Start and Comparison Group Samples

	Even Start N = 607		Comparison N = 607	
	Count	Valid %	Count	Valid %
Gender				
Female	299	49%	293	48%
Male	308	51%	314	52%
Race				
African American	45	7%	44	7%
Asian or Pacific Islander	80	13%	82	14%
Hispanic or Latino	367	60%	367	60%
Native American	2	0.3%	1	0.2%
White	113	19%	113	19%
Multi-racial	0	0%	0	0%
Low-Income Status				
Not eligible for free/reduced lunch	139	23%	139	23%
Eligible for free/reduced lunch	468	77%	468	77%
First Language Not English				
Not FLNE	169	28%	169	28%
First language not English	438	72%	438	72%
Limited English Proficiency				
Not LEP	331	55%	331	55%
LEP	276	45%	276	45%
Special Education				
Not Special Education	546	90%	546	90%
Full Inclusion	28	5%	28	5%
Partial Inclusion	24	4%	22	4%
Substantially Separate	2	0.3%	5	1%
Placed out of District	1	0.2%	0	0%
3 or 4 year olds (SY02 and SY03 only)	6	1%	6	1%
Grade at the Time of Match				
Pre-Kindergarten	51	8%	51	8%
Part-time Kindergarten	65	11%	65	11%
Full-time Kindergarten	183	30%	183	30%
Grade 1	129	21%	129	21%
Grade 2	97	16%	97	16%
Grade 3 or above	82	14%	82	14%

While the comparison group provides statistical controls for many observed demographic characteristics, the non-experimental nature of the study in light of the voluntary nature of the program leaves open the possibility for some bias, known as self-selection bias. That is, it is possible that those who choose to enroll in an intensive program, such as Even Start, may possess different characteristics than those who choose not to enroll in the program, including the family's emphasis on education and parents' motivation to help their children succeed in school. Because these characteristics cannot be readily observed and measured, it is impossible to control for them directly in a non-experimental study.

The only way to eliminate the possibility of self-selection bias is to use an experimental design whereby individuals are randomly assigned to treatment (Even Start) and control (non-Even Start) groups, or to mimic an experimental design by exploiting differences in program participation occurring on a random basis. An example would be comparing students in the program to those on waiting lists for the program, although this was not feasible for this study.

Analysis of Data

Several analyses were conducted to identify the potential effect of Even Start program participation on subsequent MCAS achievement, described in detail below. For all analyses, statistical tests were performed to identify whether any observed differences were statistically significant. Statistical significance refers to the likelihood that any variation in scores can be attributed to real difference in performance as opposed to random variation in the sample. For all analyses in this report, observed differences were considered significant at the 95% confidence level or above.

Analysis of MCAS Performance Levels of Even Start and Comparison Group Students

To maximize the statistical power of the analysis, observations were pooled across years. The proportion of students achieving scores at each of the four performance levels—*Advanced* (grade 4 ELA and Mathematics) or *Above Proficient* (grade 3 ELA); *Proficient*; *Needs Improvement*; and *Warning*—were compared for Even Start and comparison group students and chi-squared statistics were used to determine whether any observed differences were statistically significant.

Analysis of MCAS Scores of Even Start and Comparison Group Students

A second analysis compared students' MCAS scores to test for differences that might not be detectable in the analysis of score distributions, but might be observed through a more refined analysis. Because the sample was pooled, MCAS raw scores for each grade and subject test needed to be standardized by year to have a mean of zero and a standard deviation of one.⁵ This process involved adjusting individual students' raw scores by subtracting the mean raw score of all students for that year and test and dividing the result by the standard deviation—a measure of the typical variation, or average differences from the mean, in the sample. This conversion assigned students who achieved a raw score that was above average an adjusted, or standard, score greater than zero, while students who achieved a below average score were assigned a standard score less than zero.

As an example, Table 4 presents sample calculations for a subset of students who took the grade 3 ELA MCAS in 2008, which had a mean raw score of approximately 34.8 and a standard deviation of approximately 8.2. One of

⁵ The reason for this adjustment is that the distribution of raw scores and the associated scaled scores and performance levels may vary across years for individual tests. Standardizing scores adjusts for this variation. Although scaled scores provide a measure that has a consistent interpretation across years, the way in which scales are developed makes it inappropriate for use in calculating statistics and measures of variance. For more information on this, please see: Massachusetts Department of Elementary and Secondary Education. (2008). *Researcher's Guide to Massachusetts State Education Data*. Office of Strategic Planning, Research, and Evaluation. Unpublished document.

the sample students performed at about the mean (Student 3), two performed above the mean (Students 1 and 2), and two performed below the mean (Students 4 and 5). The student's performance relative to the mean is an intermediate calculation whereby the mean for that test (34.8) is subtracted from the student's actual observed raw score, while the student's standard raw score is their performance relative to the mean divided by the standard deviation for that test (8.2).

As the table illustrates, the standard raw score can essentially be interpreted as a measure of how an individual student's score varied from the statewide mean on that test, in standard deviation units. Similarly, the mean standardized score for the sample or a subgroup of the sample can be interpreted as a measure of that group's average performance relative to that of all students statewide. To provide some context, for students receiving scores in the *Needs Improvement* or *Proficient* ranges on an MCAS exam, a standard deviation is roughly equivalent to one performance level. Said another way, if a program causes a student scoring at the *Needs Improvement* level to improve his or her MCAS performance by one standard deviation, that student would move to the *Proficient* level.

**Table 4: Sample Standardization Calculations for a Subset of Students
Grade 3 ELA, Test Year = 2008 (Mean = 34.8; Standard Deviation = 8.2)**

	Student's Raw Score	Student's Raw Score Relative to the Statewide Mean (34.8)	Student's Standard Raw Score
Student 1	48	+13.2	1.6
Student 2	37	+2.2	0.3
Student 3	35	0.2	0.0
Student 4	33	-1.8	-0.2
Student 5	22	-12.8	-1.6

Both the mean raw scores and standard raw scores were compared for Even Start and non-Even Start students, but consistent with ESE's *Researcher's Guide to Massachusetts State Education Data*, t-tests to identify the statistical significance of differences were conducted using standard raw scores. In addition, effect sizes were calculated for each observed difference in standard raw scores. Unlike statistical significance tests, effect sizes are not as sensitive to sample size, but instead report the size of the observed difference between the treatment group (Even Start) and non-treatment group (comparison group) relative to the standard deviation for both groups of students. Thus, it provides a measure of the relative impact of the program, in terms of percentage of a standard deviation explained by the program.

Interpreting the magnitude of effect sizes is not an exact science. An observed effect size of 0.0 would mean that the intervention or program accounts for none of the typical variation and is having essentially no impact. The larger the effect size the more variation is explained by the program, and the more powerful the intervention. One commonly accepted practical interpretation considers effect sizes "low" when they are around 20% to 30%, "medium" at around the 50% level, and "large" when they reach 80% and beyond.⁶ However, some researchers note that in some fields where individual variation is high and many factors influence outcomes, effect sizes that reach the 20% to 30% level may be particularly difficult to reach, and using the common interpretation of effect sizes can be "misleading."⁷

⁶ Cohen, J. *Statistical power for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

⁷ For a complete discussion of this, please see: Valentine, J. C. & Cooper, H. (2003). Effect size substantive interpretation guidelines: Issues in the interpretation of effect sizes. Washington, DC: What Works Clearinghouse.

Analysis of MCAS Scores of Even Start and Comparison Group Students, by Program Category

Using a methodology consistent with the analysis just discussed, this study also compared the MCAS scores of Even Start students relative to their comparison group for subsets of local Even Start programs. Because local programs may have had different program delivery strategies and levels of implementation, it is possible that different impacts on student MCAS performance may be observed among different types of programs. Although measures regarding program features or implementation quality were not readily available for this study, federal funding for the Even Start program was dramatically reduced in 2007, causing 18 of the 22 Even Start programs operating at the time to be eliminated. ESE program managers note that the program's past performance and level of implementation were key criteria considered when determining which programs would continue to receive funding.⁸ From a research perspective this provided an opportunity to categorize programs and analyze the differences between Even Start and comparison group students in the subset of programs considered well-implemented from the perspective of ACLS.

Comparisons among Even Start Students

Finally, the study used correlation analysis and analysis of variance to identify differences among Even Start students based on details related to program participation. The program details able to be included in this study were (1) age at entry into the program and (2) duration of time in the program. These measures were calculated from the ESE-provided SMARTT databases. Age of entry into the programs represents students' chronological age in years at the time of their first reported month in Even Start. Duration of time in the program represents the time, in months, between a student's first reported enrollment in Even Start and his or her withdrawal from the program.

Correlation analysis identifies the extent to which a relationship exists between two variables. The outcome statistic of this analysis, the correlation coefficient, ranges from -1.0 to 1.0. Positive correlations between measures indicate that as one measure increases, the other also increases; negative correlations indicate that as one measure increase, the other decreases, and vice versa. A correlation at or near zero indicates no relationship was found, and the strength of the relationship is shown in the distance from 0, with the strength increasing as a correlation approaches either 1.0 or -1.0.

⁸ Other factors, including the age of the target populations and length of funding, were considered, such that this is not a precise measure. Programs exclusively serving infants and toddlers were not considered for continued funding, as were those that had been funded for 13 years or more.

Findings

This section presents the results of analyses performed to identify the potential effect of Even Start program participation on subsequent MCAS performance. Findings are based on students who participated in an Even Start program between SY02 and SY08 who could be matched to a student identifier in the SIMS data files and a valid grade 3 or grade 4 MCAS record in the 2002 through 2008 MCAS data files. Analyses include:

- Comparison of the MCAS performance level results of Even Start students to those of a matched comparison group.
- Comparison of the mean MCAS scores of Even Start students to those of a matched comparison group.
- Analysis of the MCAS score results of Even Start students relative to a matched comparison group for the four currently operating Even Start programs.
- A correlation and subgroup analyses of Even Start students by age of enrollment and duration of participation to identify if these aspects of program participation influence results.

Statistical tests were performed on all results to identify whether observed differences were statistically significant at the 95% confidence level. Although not the focus of this study, additional analyses, including performance level frequency distributions for Even Start and comparison group students on the grade 5 through grade 7 MCAS tests, are shown in the appendix to this report.

MCAS Performance Levels for Even Start and Comparison Group

The proportion of Even Start and comparison group students performing at each of the four MCAS performance levels is shown in Table 5. Overall, the performance of Even Start students did not differ substantially from that of their comparison group on either of the ELA exams. On the grade 4 Mathematics exam, a slightly larger proportion of Even Start students performed at the advanced or proficient level relative to the comparison group (27% compared to 23%), although the difference in distribution is not statistically significant, and thus, could be the result of random variation in the samples and not true differences in performance between the two groups.

Table 5: MCAS Performance Levels for Even Start and Comparison Group

	Even Start		Comparison		Total	
	Count	%	Count	%	Count	%
Grade 3 ELA						
Above Proficient	20	3%	31	5%	51	4%
Proficient	190	33%	176	30%	366	31%
Needs Improvement	274	47%	275	47%	549	47%
Warning	97	17%	105	18%	202	17%
Total	581	100%	587	100%	1168	100%
Grade 4 ELA						
Advanced	8	2%	9	2%	17	2%
Proficient	102	25%	101	25%	203	25%
Needs Improvement	216	53%	220	53%	436	53%
Warning	81	20%	82	20%	163	20%
Total	407	100%	412	100%	819	100%
Grade 4 Mathematics						
Advanced	31	8%	29	7%	60	7%
Proficient	78	19%	68	16%	146	18%
Needs Improvement	212	52%	232	56%	444	54%
Warning	87	21%	85	21%	172	21%
Total	408	100%	414	100%	822	100%

Analysis of MCAS Raw Scores and Standard Raw Scores

As mentioned previously, a more sensitive analysis involves comparing the mean scores of Even Start students on each of the tests to those of students in the comparison group. This required that MCAS raw scores for each grade and subject test be standardized by year to have a mean of zero and a standard deviation of one.⁹ This process involved standardizing individual students' raw scores by subtracting the mean raw score of all students for that year and that test from each individual score, and dividing the result by the standard deviation—a measure of the typical variation, or average differences from the mean, in the sample. This standard score can essentially be interpreted as a measure of how an individual student's score varied from the statewide mean on that test, in standard deviation units.¹⁰ For the group as a whole, the mean standard score represents a gap in performance between that group and all students who took the test over the period. (For a more complete discussion of this conversion and its interpretation, please see the Analysis of Data section on page 9 of this report.)

The results of this score analysis are displayed in Table 6. As this table shows, Even Start students received slightly higher mean raw scores on all tests than those in the comparison groups (0.2 additional raw score points on the grade 3 ELA test, 0.3 additional raw score points on the grade 4 ELA test, and 0.5 points on the grade 4 Mathematics test).

⁹ The reason for this standardization is that the distribution of raw scores and the associated scaled scores and performance levels may vary across years for individual tests. Standardizing scores adjusts for this variation. A more detailed discussion of this adjustment is presented in the methodology section of this report.

¹⁰ For the purposes of comparison, for students at the *Needs Improvement* and *Proficient* levels, a standard deviation is roughly equivalent to a performance level.

The same trend can be seen in the pattern of standard raw scores, on which statistical tests were able to be performed. The first thing to note is that because the standard scores express student performance *relative to statewide averages*, negative scores for both Even Start and comparison group students indicate that both populations received raw scores that were lower than the mean raw score for all students in the year they were tested. This is not unexpected given the difficulty of the population Even Start serves. For example, the mean standard raw score of -0.555 for Even Start students on the grade 3 ELA exam indicates that, on average, Even Start students performed approximately 0.555 standard deviations less well (i.e., a gap in performance) on that test than the statewide mean in the year they were tested. On that same test, comparison group students received raw scores that were, on average, 0.592 standard deviations lower than the mean. Therefore, although Even Start students received scores that were lower on average than all tested students statewide, they performed slightly better relative to the comparison group.

Overall, the mean standard raw scores on all three tests show Even Start students performing slightly better than the comparison group of students, such that the gap in achievement relative to the statewide mean is lower for Even Start students. However, differences were small and not statistically significant.¹¹

Table 6: Mean MCAS Raw Scores and Standard Raw Scores for Even Start and Comparison

	Count	Mean Raw Score	Standard Raw Score	
			Mean	Std Dev.
Grade 3 ELA				
Even Start	581	30.1	-0.555	1.098
Comparison	587	29.9	-0.592	1.109
Difference	n/a	0.2	0.037	-
Grade 4 ELA				
Even Start	407	45.5	-0.552	1.002
Comparison	412	45.1	-0.589	1.016
Difference	n/a	0.3	0.038	-
Grade 4 Mathematics				
Even Start	408	32.8	-0.430	1.010
Comparison	414	32.4	-0.479	0.972
Difference	n/a	0.5	0.050	-

Effect sizes, which report the percent of a standard deviation unit explained by the treatment, were also generally small, ranging from approximately 3% on the grade 3 ELA exam to approximately 5% on the grade 4 Mathematics exam. These would generally be considered *extremely small* impact effect sizes, as only a small percent of overall variation can be explained by differences in the treatment.¹²

¹¹ Statistical significance refers to whether the variation in scores is attributed to random variation within the sample or to a real difference in performance. In this case, the variation can only be attributed to random variation.

¹² Effect sizes are generally considered small when they approach 0.20 or 20% of a standard deviation, although where substantial individual variation exists and many factors influence outcomes, as is the case in education, effect sizes may be expected to be smaller. See Valentine, J. C. & Cooper, H. (2003). Effect size substantive interpretation guidelines: Issues in the interpretation of effect sizes. Washington, DC: What Works Clearinghouse.

Analysis of MCAS Outcomes of Students from Currently Operating Programs

Table 7 presents the results of the analysis of the performance of students who attended one of the four programs that continued to received funding after 2007—Lowell Public Schools, YMCA of Greater Boston/Otis School, Waltham Public Schools, and YMCA of Greater Fall River—and were considered well-implemented Even Start comprehensive family literacy programs by program managers. Among this subset of programs, differences between Even Start and comparison group students were larger than for all programs as a whole. For example, on the grade 3 ELA test, Even Start students performed 0.9 raw score points, or 0.125 standard deviations better than comparison group students. Among all programs, the difference was only 0.2 raw score points or 0.037 standard deviations. These differences were not statistically significant.

Table 7: Mean MCAS Raw Scores and Standard Raw Scores for Even Start and Comparison, Currently Operating Even Start Programs Only

	Count	Mean Raw Score	Standard Raw Score	
			Mean	Std Dev.
Grade 3 ELA				
Even Start	124	30.0	-0.532	1.034
Comparison	122	29.1	-0.657	1.173
Difference	n/a	0.9	0.125	-
Grade 4 ELA				
Even Start	82	46.0	-0.491	0.943
Comparison	87	44.5	-0.636	0.932
Difference	n/a	1.5	0.144	-
Grade 4 Mathematics				
Even Start	82	32.7	-0.428	0.924
Comparison	87	32.1	-0.495	0.910
Total	n/a	0.6	0.066	-

Effect sizes observed for these four programs were generally larger than for all Even Start programs as a whole, and were 11% and 15% on the grade 3 and grade 4 ELA exams, respectively, and 7% for the grade 4 Mathematics exam. This provides some evidence that well-implemented comprehensive family literacy programs can, in fact, have a small impact on student's subsequent MCAS achievement, particularly in ELA.

Table 8 presents the results of an analysis of the performance of students who had attended one of the programs eliminated in 2007 relative to their comparison group. Programs in this tier include: Attleboro Literacy Center, Cambridge Public Schools, Clinton Public Schools, Community Action North Quabbin, Greater Lawrence Community Action Council, Haverhill Public Schools, Holyoke Public Schools, Lynn Public Schools, Malden Public Schools, Northampton Public Schools, P.A.C.E./New Bedford, Pittsfield Public Schools, Watertown Public Schools, and Worcester State College/Latino Education Institute.¹³ It should be noted that because funding was extremely limited and other criteria were considered in the funding decisions, not all programs in the tier may have been considered to be less well implemented than those that had their funding continued. As a result, the classification may be somewhat imprecise, hence limiting the statistical rationale for the analysis. On average, students attending one of these programs actually performed slightly less well than a comparison group of students, although the differences were small and not statistically significant.

¹³ Some discontinued schools were not included because none of their students had taken the MCAS test during the study period. This affected primarily programs targeting infants and toddlers.

Table 8: Mean MCAS Raw Scores and Standard Raw Scores for Even Start and Comparison, Programs Eliminated in 2007

	Count	Mean Raw Score	Standard Raw Score	
			Mean	Std Dev.
Grade 3 ELA				
Even Start	335	29.8	-0.619	1.127
Comparison	343	30.0	-0.601	1.066
Difference	n/a	-0.2	-0.018	-
Grade 4 ELA				
Even Start	223	44.5	-0.662	1.007
Comparison	227	44.8	-0.635	1.080
Difference	n/a	-0.3	-0.027	-
Grade 4 Mathematics				
Even Start	225	31.6	-0.571	1.011
Comparison	226	31.9	-0.535	0.953
Total	n/a	-0.3	-0.036	-

Finally, Table 9 presents the performance of Even Start students from the remaining programs, which were discontinued prior to the budget cuts of 2007. Programs in this tier include: Community Action Network of Amesbury/Newburyport, Narragansett, North Adams, Randolph, Revere, Hurley School in Boston, and the Vietnamese American Civic Association, also in Boston. These programs were not eliminated at the same time and for the same reasons, limiting the statistical rationale for considering them a single subgroup.

Overall, this group of students tended to outperform their non-Even Start counterparts on all three tests, and differences on the grade 4 Mathematics test were the largest for students in this subgroup of programs relative to the others. Again, none of these results were found to be statistically significant.

Table 9: Mean MCAS Raw Scores and Standard Raw Scores for Even Start and Comparison, Programs Eliminated Prior to 2007

	Count	Mean Raw Score	Standard Raw Score	
			Mean	Std Dev.
Grade 3 ELA				
Even Start	122	31.1	-0.405	1.076
Comparison	122	30.4	-0.501	1.163
Difference	n/a	0.7	0.097	-
Grade 4 ELA				
Even Start	102	47.3	-0.359	1.014
Comparison	98	46.5	-0.443	0.926
Difference	n/a	0.8	0.083	-
Grade 4 Mathematics				
Even Start	101	35.7	-0.116	1.013
Comparison	101	33.5	-0.342	1.058
Total	n/a	2.1	0.225	-

MCAS Results by Age at Program Enrollment and Total Months Attended

Finally, this study used correlation analyses to identify whether age at program entry and duration of time in the program were related to student MCAS performance. Correlation analysis compares two variables and reports a single value that indicates whether the variables tend to move together, and if so, in what direction. Correlations can be positive (as one variable increases, the other does as well) or negative (as one variable increases, the other decreases and vice versa). Correlation coefficients range from -1.0 to 1.0, and the closer the observed coefficient to either -1.0 (negative relationship) or 1.0 (positive relationship), the stronger the relationship. Coefficients observed here were generally relatively weak, although this would not be unexpected given the number of factors that may influence students' MCAS performance.

Table 10 presents the results of the correlation analysis regarding student age at program entry and their subsequent standard raw scores on the MCAS. The negative coefficients indicate that the younger a student was at the time of program entry, the better the student performed, on average, on their subsequent MCAS tests relative to other students who took that same test. Observed coefficients were statistically significant for both ELA exams, but not for the Mathematics exam. Coefficients were all relatively small, showing that age at program entry is only a weak predictor of subsequent MCAS performance.

**Table 10: Correlation Coefficients
Age at Program Entry and Subsequent MCAS Performance (Standard Raw Scores)
Even Start Students Only**

Test	Correlation Coefficient
Grade 3 ELA	-0.163
Grade 4 ELA	-0.160
Grade 4 Mathematics	-0.082

This relationship can also be seen in a subgroup analysis of students' MCAS scores by age at program entry, as shown below in Table 11. Students were categorized into three groups, those who entered the program at age four or younger, those who were between five and six when they entered the program, and those who were seven and older. On each of the three MCAS tests, students who were younger at the time of their enrollment in Even Start tended to outperform students who enrolled when they were older. Based on analysis of variance and Scheffe post hoc tests, there is a statistically significant difference between ages four or younger and both older groups on the grade 3 ELA test, and between ages four or younger and ages 7 and older on the grade 4 ELA test. The differences on the grade 4 Mathematics test are not statistically significant.

**Table 11: Mean MCAS Raw Scores and Standard Raw Scores
Even Start Students, by Age at Program Entry**

	Count	Mean Raw Score	Standard Raw Score	
			Mean	Std. Dev.
Grade 3 ELA				
Ages 4 and under	164	32.7	-0.335	1.094
Ages 5 to 6	237	30.5	-0.571	1.047
Ages 7 and over	180	27.3	-0.736	1.138
Grade 4 ELA				
Ages 4 and under	65	47.9	-0.357	0.978
Ages 5 to 6	169	46.4	-0.481	0.928
Ages 7 and over	173	43.7	-0.694	1.065
Grade 4 Mathematics				
Ages 4 and under	65	34.1	-0.308	1.073
Ages 5 to 6	170	33.6	-0.393	0.938
Ages 7 and over	173	31.6	-0.511	1.051

Finally, Table 12 shows the results of correlation analysis of duration of program participation, in months, and students' performance on their subsequent MCAS exams. Overall, this analysis showed that time in the program did not correlate with student performance on either of the ELA exam, as indicated by the near zero correlation coefficients for each of these two tests. However, a weak statistically significant positive correlation was found between months attended and student performance on the grade 4 Mathematics exam (0.112). That is, students who spent more time in Even Start performed slightly better than those who spent less time in the program.

**Table 12: Correlation Coefficients
Duration of Program Participation and Subsequent MCAS Performance (Standard Raw Scores)
Even Start Students Only**

Test	Correlation Coefficient
Grade 3 ELA	0.036
Grade 4 ELA	0.059
Grade 4 Mathematics	0.112

Appendix A: Comparison to Statewide Low-Income and LEP Students

At ESE's request, the MCAS performance of Even Start students was compared to that of all low-income and limited English proficient (LEP) or formerly limited English proficient (FLEP) students tested statewide during the study period. These students serve as potentially relevant comparison groups because Even Start program requirements and design features make it likely that participating students were members of one or both groups. In fact:

- Program managers note that because Even Start serves students in high need families, nearly all participating students are classified as low-income upon entry in the program. Data on the low-income status of students at the time of program participation was not available for this study.
- An analysis of SMARTT data reveals that nearly two-thirds of all students participating in Even Start between SY02 and SY08 were from families whose native language was not English. However, of those who could be matched to a grade 3 or grade 4 MCAS record—and thus included in the sample—approximately 45% were considered LEP at the time of program enrollment or their first appearance in SIMS, whichever came first.

While these comparisons may be useful, the inability to control for district characteristics and other factors that may influence student achievement but are more or less prevalent among Even Start students serve as critical limitations to findings drawn from the analyses. As such, these comparisons may be particularly subject to bias, both as a result of the lack of statistical controls for other observed characteristics and as a result of participants' self-selection into the program.

For these analyses, standardizing students' raw scores is even more critical, given that the proportion of students taking the test in each year differed for Even Start students and all test takers, and by extension, comparison low-income and LEP students. As an example, Table 13 presents the proportion of grade 3 students taking the test each year, showing that a much larger proportion of Even Start students took the test in the later years. Because overall mean scores on the MCAS improved over this period, simply looking at raw scores could be misleading, as there are more Even Start students, proportionally, from the higher-performing later years. In contrast, the matched comparison sample, by virtue of matching students in the same year, displayed similar proportions of Even Start and comparison group students taking the test in each year.

**Table 13: Proportion of Students Taking the Grade 3 MCAS Each Year
Even Start Sample and All Students**

Year	Even Start	All Students
2002	2%	15%
2003	4%	15%
2004	9%	14%
2005	12%	14%
2006	17%	14%
2007	27%	14%
2008	30%	14%
Total	100%	100%

Table 14 compares the mean scores of Even Start students with all low-income students statewide. The negative standard raw scores, which can be interpreted as an average gap between that group's students and the statewide average for all students, reveal that on all three tests, both Even Start students and all low-income students performed below the statewide mean. However, students participating in the Even Start program tended to outperform low-income students, on average, on all tests, but the difference was statistically significant only with respect to the grade 4 Mathematics exam. On that test, the gap in performance relative to the statewide mean for Even Start students was 0.430 standard deviations, which is smaller than the performance gap for all low-income students (0.549 standard deviations).

**Table 14: Mean MCAS Raw Scores and Standard Raw Scores
Even Start and Statewide Low-Income Students**

	Count	Mean Raw Score	Standard Raw Score	
			Mean	Std Dev.
Grade 3 ELA				
Even Start	581	30.1	-0.555	1.098
All Low Income	148,399	28.8	-0.565	1.096
Difference	n/a	1.3	0.010	-
Grade 4 ELA				
Even Start	407	45.5	-0.552	1.002
All Low Income	150,501	44.8	-0.586	1.061
Difference	n/a	0.7	0.034	-
Grade 4 Mathematics				
Even Start	408	32.8	-0.430	1.010
All Low Income	151,406	30.1	-0.549	1.021
Difference	n/a	2.7	0.119	-

Table 15 compares the MCAS performance of a subset of Even Start students who were either LEP or FLEP to that of all LEP and FLEP students tested during the study period. The Even Start group was limited to students who were LEP or FLEP because not all Even Start students were from families that were non-native English speakers.

The large negative standard raw scores for statewide LEP and FLEP students indicate that statewide, students from this subgroup performed substantially lower than the mean for the entire population of students taking the test. Performance gaps ranged from 0.653 standard deviations on the grade 4 Mathematics exam to more than 0.84 standard deviations on the grade 3 and grade 4 ELA exams. Given that a standard deviation is approximately equal to a performance level for students performing at *Needs Improvement* and *Proficient* levels, these gaps represent sizeable achievement differences for LEP and FLEP students relative to statewide averages.

LEP and FLEP students who participated in Even Start also performed below statewide averages, but fared substantially better on all tests statewide than LEP and FLEP students as a whole. That is, the performance gaps were much lower than those observed for LEP students as a whole. For example, on the grade 4 ELA exam, Even Start LEP and FLEP students performed 0.602 standard deviations below the average for all test takers, but this was substantially better than the performance of all Even Start students. Differences on the grade 4 tests were statistically significant, while the difference on the grade 3 test was not.

**Table 15: Mean MCAS Raw Scores and Standard Raw Scores
Limited English Proficient and Formerly Limited English Proficient Students, Even Start and Statewide**

	Count	Mean Raw Score	Standard Raw Score	
			Mean	Std. Dev.
Grade 3 ELA				
Even Start LEP and FLEP	320	28.5	-0.730	1.116
Statewide LEP and FLEP	38,007	27.1	-0.842	1.175
Difference	n/a	1.4	0.112	-
Grade 4 ELA				
Even Start LEP and FLEP	225	45.0	-0.602	1.005
Statewide LEP and FLEP	35,350	42.3	-0.841	1.175
Difference	n/a	2.7	0.238	-
Grade 4 Mathematics				
Even Start LEP and FLEP	226	32.6	-0.455	1.008
Statewide LEP and FLEP	35,600	29.9	-0.653	1.114
Difference	n/a	2.7	0.198	-

Appendix B: Additional MCAS Grade Levels

The following two tables present additional analyses showing the proportion of students in the Even Start and comparison groups performing at each MCAS performance level on the grade 5 through grade 7 ELA and Mathematics exams. Grade 8 and grade 10 are excluded because the sample sizes were insufficient for reporting.

Table 16 presents results on the grade 5 through grade 7 ELA exams. Slightly larger proportions of Even Start students scored in the *Advanced* and *Proficient* ranges relative to the comparison group on the grade 5 (39% compared with 36%) and the grade 6 exams (45% compared with 42%), which is generally consistent with the analysis of the grade 3 and grade 4 ELA results presented in Table 5 of this report. However, on the grade 7 exam, Even Start students were less likely to receive scores in the *Advanced* and *Proficient* ranges than were comparison group students (48% compared with 58%). However, none of the differences in performance patterns were statistically significant, and as such, they cannot be attributed to true differences as opposed to random variation.

Table 16: ELA MCAS Performance Levels of Students on Grade 5 through Grade 7 Exams

	Even Start		Comparison		Total	
	N	%	N	%	N	%
Grade 5 English Language Arts						
Advanced	11	5%	7	3%	18	4%
Proficient	70	34%	72	33%	142	33%
Needs Improvement	98	47%	105	48%	203	48%
Warning	29	14%	34	16%	63	15%
Total	208	100%	218	100%	426	100%
Grade 6 English Language Arts						
Advanced	10	8%	3	2%	13	5%
Proficient	46	37%	50	40%	96	38%
Needs Improvement	52	41%	56	44%	108	43%
Warning	18	14%	17	13%	35	14%
Total	126	100%	126	100%	252	100%
Grade 7 English Language Arts						
Advanced	3	4%	2	3%	5	3%
Proficient	33	44%	42	55%	75	49%
Needs Improvement	30	40%	26	34%	56	37%
Warning	9	12%	7	9%	16	11%
Total	75	100%	77	100%	152	100%

Table 17 presents results on the grade 5 through grade 7 Mathematics exams. On each of these three tests, Even Start students were more likely to receive scores in the *Advanced* and *Proficient* range and less likely to receive scores in the *Warning* range, although none of the differences was statistically significant. This is generally consistent with the analysis of the grade 4 Mathematics results presented in Table 5. The differences were most notable at the grade 7 level, where 37% of Even Start students received scores in the *Advanced* or *Proficient* levels, compared to 29% of comparison group students.

Table 17: Mathematics MCAS Performance Levels of Students on Grade 5 through Grade 7 Exams

	Even Start		Comparison		Total	
	N	%	N	%	N	%
Grade 5 Mathematics						
Advanced	10	5%	17	8%	27	6%
Proficient	55	27%	49	23%	104	25%
Needs Improvement	81	39%	86	40%	167	39%
Warning	60	29%	65	30%	125	30%
Total	206	100%	217	100%	423	100%
Grade 6 Mathematics						
Advanced	18	13%	10	7%	28	10%
Proficient	33	23%	36	26%	69	24%
Needs Improvement	47	33%	48	34%	95	34%
Warning	44	31%	46	33%	90	32%
Total	142	100%	140	100%	282	100%
Grade 7 Mathematics						
Advanced	5	7%	7	9%	12	8%
Proficient	22	30%	15	20%	37	25%
Needs Improvement	24	33%	22	30%	46	31%
Warning	22	30%	30	41%	52	35%
Total	73	100%	74	100%	147	100%