MASSACHUSETTS DEPARTMENT OF ELEMENTARY AND SECONDARY EDUCATION

Massachusetts Early Literacy Tutoring Study

Final Report

Daniel Hubbard, Katherine Gross, Aleksandra Holod, Emma Wilson, Mark Lachowicz, Emily Ackman

June 2022

American Institutes for Research® | AIR.ORG

AIR®
Advancing Evidence. Improving Lives.

Massachusetts Early Literacy Tutoring Study

Final Report

Daniel Hubbard, Katherine Gross, Aleksandra Holod, Emma Wilson, Mark Lachowicz, Emily Ackman

June 2022

AIR® Headquarters
1400 Crystal Drive, 10th Floor
Arlington, VA 22202-3289
+1.202.403.5000 | **AIR.ORG**

18108_06/22

Contents

Executive Summary 1

Introduction 2
- Overview of Early Literacy Tutoring 2
- Overview of the Evaluation 2

Summary of Services 6
- Tutoring Services Delivered 6
- Students Served 8
- Catapult 10
- Literacy Lab 14
- Springboard 18

Summary of Student Learning Outcomes 22
- Associations Between Program Characteristics and Student Outcomes 24
- Associations Between Student Characteristics and Student Outcomes 27
- What Works for Whom, and Under Which Conditions? 28
- Cost Analysis 29

Limitations 31

Conclusion 32

References 33

Appendix A. Variable Sources and Measurement Properties 34

Exhibits

Exhibit 1. Research Questions and Data Sources 3

Exhibit 2. Survey Sample Sizes and Response Rates 5

Exhibit 3. Snapshot of Tutoring Delivered Across the Three Vendors 7

Exhibit 4. Home Districts of the Students Served Across the Three Vendors 7

Exhibit 5. Grade Levels of the Students Served Across the Three Vendors 8

Exhibit 6. Baseline DIBELS Reading Levels of the Students Served Across the Three Vendors 9

Exhibit 7. Demographic Snapshot of the Students Served Across the Three Vendors 10

Exhibit 8. Snapshot of Catapult Tutoring Services included in this analysis 10

Exhibit 9. Home Districts of the Students Served by Catapult 12

Exhibit 10. Grade Levels of the Students Served by Catapult 12

Exhibit 11. Baseline DIBELS Reading Levels of the Students Served 13

Exhibit 12. Demographic Snapshot of the Students Served by Catapult 14

Exhibit 13. Snapshot of Literacy Lab Tutoring Services included in this analysis 15

Exhibit 14. Home Districts of the Students Served by Literacy Lab 16

Exhibit 15. Grade Levels of the Students Served by Literacy Lab 16

Exhibit 16. Baseline DIBELS Reading Levels of the Students Served by Literacy Lab 17

Exhibit 17. Demographic Snapshot of the Students Served by Literacy Lab 18

Exhibit 18. Snapshot of Springboard Tutoring Services included in this analysis 19

Exhibit 19. Home Districts of the Students Served by Springboard 20

Exhibit 20. Grade Levels of the Students Served by Springboard 20

Exhibit 21. Baseline DIBELS Reading Scores of the Students Served 21

Exhibit 22. Demographic Snapshot of the Students Served by Springboard 22

Exhibit 23. Average Student Growth on DIBELS Composite Scores 23

Exhibit 24. Average Student Growth on DIBELS Scores, by Subscale 24
Exhibit 25. Vendor-Specific Effects on Student Outcomes 25
Exhibit 26. Relationships Between Program Factors and Student Growth 26
Exhibit 27. Relationships Between Student Characteristics and Student Growth 27
Exhibit 28. Estimated Replication Cost Breakdowns, by Vendor 30
Exhibit 29. Summary of Replication Cost and Cost-Effectiveness Findings 31

Executive Summary

This summative evaluation of the Early Literacy Supplemental Services Tutoring program aimed to gather information on predictors of student growth, program models, and cost-effective approaches for future implementation. The information in this evaluation came from data made available to the evaluation team for the period from September 2021 to April 2022.

- Over the 2021/22 academic year, the program served an estimated 3,000 kindergarten, first, second, and third grade students across 16 school districts.
- 1,008 students who received a total of 22,932 sessions of tutoring are included in the evaluation analysis. Tutoring services were evaluated in 35 high-need elementary schools across 10 districts.
- Eighty-five percent of students evaluated received tutoring scored below benchmark on the Dynamic Indicators of Basic Early Literacy Skills (DIBELS) reading assessment at the beginning of tutoring services.
- On average, students' composite scores on the DIBELS increased by 16 points over a 3- to 4-month span, corresponding to about half of the growth expected on the DIBELS over a full school year.
- Hispanic/Latinx and Asian students showed stronger literacy gains than White or Black students who received equivalent tutoring services.
- Students with lower pretest scores showed stronger literacy gains than students with higher baseline scores, suggesting that the tutoring services overall are closing opportunity gaps.
- The cost of tutoring services ranged from $925 to $1,909 per student.

Introduction

Overview of Early Literacy Tutoring

Only 56% of Massachusetts third-graders were meeting proficiency targets for English language arts prior to the pandemic (Massachusetts Department of Elementary and Secondary Education [DESE], 2019). In addition, only 38% of the state's Black and Latinx students had reached proficiency in reading by third grade. School closures during the pandemic intensified the urgency of addressing students' literacy learning needs. The percentage of students meeting expectations decreased from 56% in 2019 to 51% in 2021 (DESE, n.d.). In response, DESE created the Early Literacy Supplemental Services Tutoring program to address the need for early literacy supports, using funding from the Governor's Emergency Education Relief (GEER) Fund.

The Early Literacy Supplemental Services Tutoring program acknowledges that remote instruction during the COVID-19-related school closures led to learning opportunity gaps. Some students now need additional supports to meet grade-level reading benchmarks. A recent meta-analysis has indicated that literacy tutoring is exceptionally effective for early elementary-grade students, and even tutoring provided by nonprofessionals can have significant positive impacts (Nickow et al., 2020). DESE engaged several tutoring vendors who implemented different tutoring program models in elementary schools across districts in Massachusetts.

Overview of the Evaluation

DESE contracted with the American Institutes for Research® (AIR®) to conduct a formative and summative evaluation of the Early Literacy Supplemental Services Tutoring program over the period from September 2021 to April 2022. This report focuses on the findings from the summative evaluation, which incorporated quantitative data sources and analytic methods to examine the learning outcomes of participating students.

About the Study

The goal of the study was to evaluate how variations in service delivery are related to the literacy outcomes of a diverse group of students, in addition to assessing the state's overall return on investment of the Early Literacy Supplemental Services Tutoring program. The study used a mixed-methods evaluation design that incorporated both quantitative and qualitative data sources and analytic methods.

AIR produced a formative evaluation in February 2022 to support improvements in the implementation and delivery of early literacy tutoring services. The formative evaluation used a qualitative design that incorporated findings from interviews and focus groups with program

stakeholders including tutoring program leaders, tutoring program coordinators who were embedded in districts or schools, principals, tutoring staff, and parents/guardians, as well as a tutor survey, student rosters, tutor rosters, and tutoring logs.

Research Questions

Exhibit 1 lists the research questions that guided the summative evaluation. To answer these questions, the study team gathered and analyzed data, including tutoring program logs, student assessments, tutor surveys, and program cost information.

Exhibit 1. Research Questions and Data Sources

Research questions	Tutoring service logs	Tutoring leader interviews	Tutor interviews	Tutor survey	State student records	DIBELS assessment	Vendor cost survey
1. Was there a significant difference between the impact of providers' program models on student outcomes?	✓	✓	✓	✓	✓	✓	
2. Which program factors best predicted student growth?	✓	✓	✓	✓		✓	
3. Were there significant differences in how students developed specific skills that were impacted by program factors?	✓	✓	✓	✓	✓	✓	
4. Which student characteristics best predicted student growth?					✓	✓	
5. Were there significant differences in how students developed specific skills that were impacted by student characteristics?	✓	✓	✓	✓	✓	✓	
6. What significant interaction effects were found between any of the variables listed above?	✓	✓	✓	✓	✓	✓	
7. What is the most cost-effective approach to future implementation and funding of early literacy programming?		✓		✓		✓	✓

Note. DIBELS is Dynamic Indicators of Basic Early Literacy Skills.

Methods

Participants

This evaluation study focuses on three tutoring programs: Catapult Learning (hereafter, Catapult), Literacy Lab, and Springboard Collaborative (hereafter, Springboard). DESE funded five vendors to provide tutoring services. However, two vendors did not participate in the evaluation.[1] AIR invited the three participating vendors to take part in the data collection activities, which the study team designed to gather data to address the research questions. The study team focused on the students served by the three tutoring vendors between September 2021 and April 2022. This observation window provides for only a snapshot of the ongoing Early Literacy Supplemental Services Tutoring program, which is currently in its second year of operation.[2]

The summary of services sample (hereafter, "students served") consists of participating students with both roster data and verifiable records of attendance in the tutoring log data. The regression sample consists of participating students with roster data, baseline (pretest) and follow-up (posttest) DIBELS assessment data, tutoring log data, and state student records. (See Appendix A for a more detailed description of sample definitions, measures, and analytic methods.)

Student Rosters, Tutor Rosters, and Tutoring Logs. All three participating vendors shared rosters and logs with the AIR study team, which allowed the team to determine how much tutoring each student received and with which tutor. Tutor rosters included the tutor names, school names, and email addresses. Student rosters included student names, demographic data, school, and district. The tutoring logs contained dates, tutor names, student names, school names, and modality (virtual/in-person), thereby providing a record of student attendance for each tutoring session.

Student Assessments. All three participating vendors shared student literacy assessments with the study team. The AIR study team trained tutoring vendors and tutors to administer the Dynamic Indicators of Basic Early Literacy Skills (DIBELS) assessment with students. Tutoring vendors gathered DIBELS and Renaissance Star Early Literacy assessment (STAR) data from tutor-administered assessments and from school district data. The student assessments were collected at the start of tutoring to provide a measure of baseline achievement and at the midyear point to show students' literacy learning during their participation in tutoring services.

[1] Tufts University chose not to participate in the evaluation. DESE discontinued the YMCA's contract to provide tutoring services.

[2] Additionally, the study team removed one district served by Springboard (Boston) and three districts served by Catapult (Chicopee, Malden, and Worcester) from all samples defined for analysis. Boston did not allow the administration of the DIBELS assessment, and Chicopee administered the Acadience assessment instead of DIBELS. Malden and Worcester launched tutoring services between February and March of 2022, leaving insufficient time to collect both baseline and follow-up DIBELS assessment data for students served within the observation window. Thus, the study team did not anticipate outcome data for any students in these four districts.

Using the DIBELS assessment data, the study team calculated baseline (pretest) and follow-up (posttest) composite scores for students. Composite scores and benchmarks are calculated according to the middle-of-year standards.[3,4]

Tutor Survey. The AIR study team administered a web-based survey to tutors, which gathered data on the following topics: tutor training, the content and structure of the tutoring sessions, assessments administered to students, differentiation of instruction, coaching and supervision of tutors, and tutors' background. Survey participation was by email, using each tutor's contact information provided in the tutor rosters. Participants received a $25 Target gift card after completing the survey. Tutors who did not complete the survey received several reminders regarding the opportunity to participate and receive the gift card. One hundred forty-three tutors completed the survey, for a response rate of 75%. Exhibit 2 presents the survey sample size and response rate for each vendor, as well as for the full study sample.

Exhibit 2. Survey Sample Sizes and Response Rates

Vendor	Surveys sent	Surveys completed	Response rate
Catapult Learning	76	49	64%
Literacy Lab	7	6	86%
Springboard	108	88	81%
All	**191**	**143**	**75%**

Analytic Approach. We conducted descriptive tabulations of the program data to determine the number of students and communities served, the demographic characteristics of participating students, and the magnitude of student learning gains. We followed this descriptive analysis with a series of regression analyses to answer questions about the associations between program factors, student demographic characteristics, and student learning outcomes. Students whose primary tutors did not complete the survey are included in the regression sample, with the survey variables set to zero and an indicator for "survey missing" set to 1. See Appendix A for more information on the variables, their sources, and their measurement properties.

Report Organization

Findings will be presented at the vendor level and then in aggregate. This will be followed by Discussion and Limitations sections.

[3] See https://dibels.uoregon.edu/sites/dibels1.uoregon.edu/files/2021-10/UO_Dibels8_Scoring_Guide_100121.pdf for more information.
[4] For cases with missing DIBELS subtest scores, the study team created a linear regression imputation model to estimate the missing scores within grade-level samples, using the available subtest scores as predictor variables. Twenty-eight students in the regression sample (15%) had pretest composite scores calculated with imputed subtest data, and 27 students in the regression sample (14%) had posttest composite scores calculated with imputed subtest data.

Summary of Services

In this section, we report the total volume of services delivered to the participant sample of 1,008 students from September 2021 to April 2022 by the three tutoring vendors who participated in the evaluation. We describe the population of the students served. Then, we present summaries for each vendor, describing their program models and the students that they served.

Tutoring Services Delivered

A participant sample of 1,008 students, in 35 schools, and 10 districts across the Commonwealth between September 2021 and April 2022, were included in the evaluation.[5] Catapult accounted for 43% of students served, with supplementary services provided at 15 schools across five districts. Literacy Lab accounted for 11% of students served, with supplementary services provided at seven schools across two districts. Last, Springboard accounted for 46% of students served, with supplementary services provided at 13 schools across three districts.

Collectively, according to tutoring log data, the three tutoring vendors provided 22,932 total tutoring sessions during the course of the study. Students participated in an average of 22.75 tutoring sessions during the study observation window, and an average of 2.22 tutoring sessions per week.

Exhibit 3 presents a snapshot of tutoring included across the three vendors, including the students and communities served, dosage, and average per-student expenditures.

[5] For the purposes of this section, the sample is defined as participating students who have been successfully matched between vendor-provided student rosters and tutoring log attendance records. See Methods for more information about how samples were defined for analysis.

Exhibit 3. Snapshot of Tutoring Included Across the Three Vendors

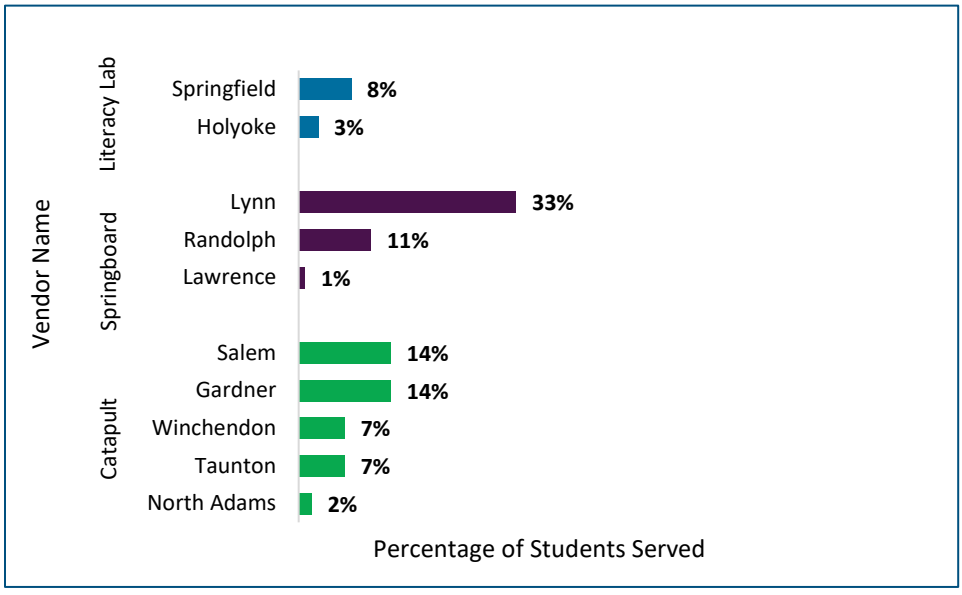
Participating Students	Schools	Districts
1,008	35	10

Number of Total Tutoring Sessions	Number of Tutoring Sessions Per Student
22,932	22.75

Tutoring Sessions Per Student Per Week	Average Expenditure Per Student
2.22	$1,051

The greatest proportion of students included in the study (33%) was enrolled in Lynn Public Schools, followed by Salem and Gardner Public Schools (14% each), Randolph Public Schools (11%), Springfield Public Schools (8%), and Taunton and Winchendon Public Schools (7% each). Holyoke Public Schools (3%), North Adams Public Schools (2%), and Lawrence Public Schools (1%) were the least represented districts among students served. Exhibit 4 presents a snapshot of the home districts of the students served, clustered by vendor.

Exhibit 4. Home Districts of the Students Served Across the Three Vendors

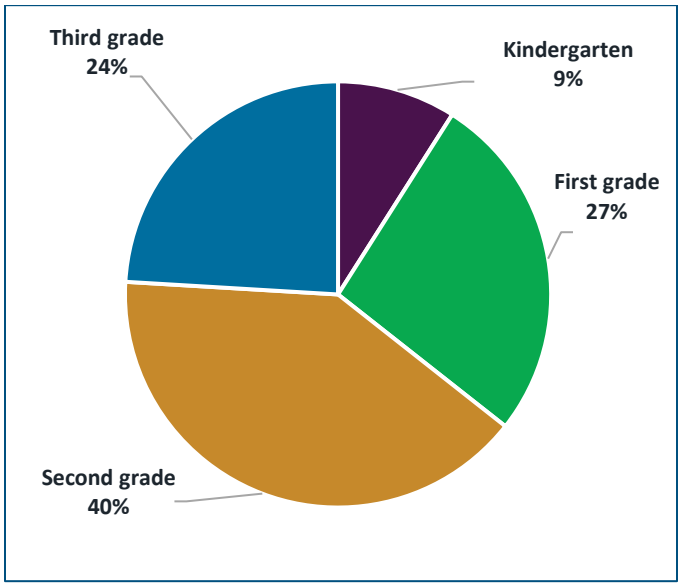


Note. N = 1,008.

Students Served

According to state student records provided by DESE, 9% of students served across the three vendors were in kindergarten, 27% were in first grade, 40% were in second grade, and 24% were in third grade. Exhibit 5 presents a snapshot of the grade levels of the students served across all three vendors.

Exhibit 5. Grade Levels of the Students Served Across the Three Vendors

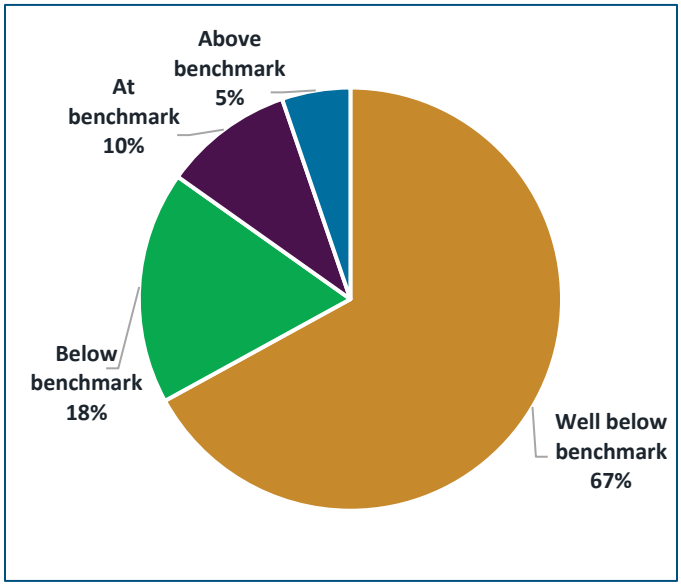


Note. N = 977. The study team used a fuzzy matching program to probabilistically match the names of the students served across the merged roster, assessment, and tutoring log data, and the state student records provided by DESE. Approximately 3% of all students served could not be matched to the state student records.

Of the students included in our analysis of learning outcomes, 67% scored well below benchmark on the DIBELS at the beginning of tutoring services, 18% scored below benchmark, 10% scored at benchmark, and 5% scored above benchmark. Students who score well below benchmark are classified as "at risk" for not meeting grade-level reading proficiency goals, whereas students who score below benchmark are classified as "some risk," students who score at benchmark are classified as "minimal risk," and students who score above benchmark are classified as "negligible risk."[6] Exhibit 6 presents the baseline DIBELS reading levels of the students served across all three vendors.

[6] See https://dibels.uoregon.edu/sites/dibels1.uoregon.edu/files/2021-06/DIBELS8thEditionGoals.pdf for more information.

Exhibit 6. Baseline DIBELS Reading Levels of the Students Served Across the Three Vendors



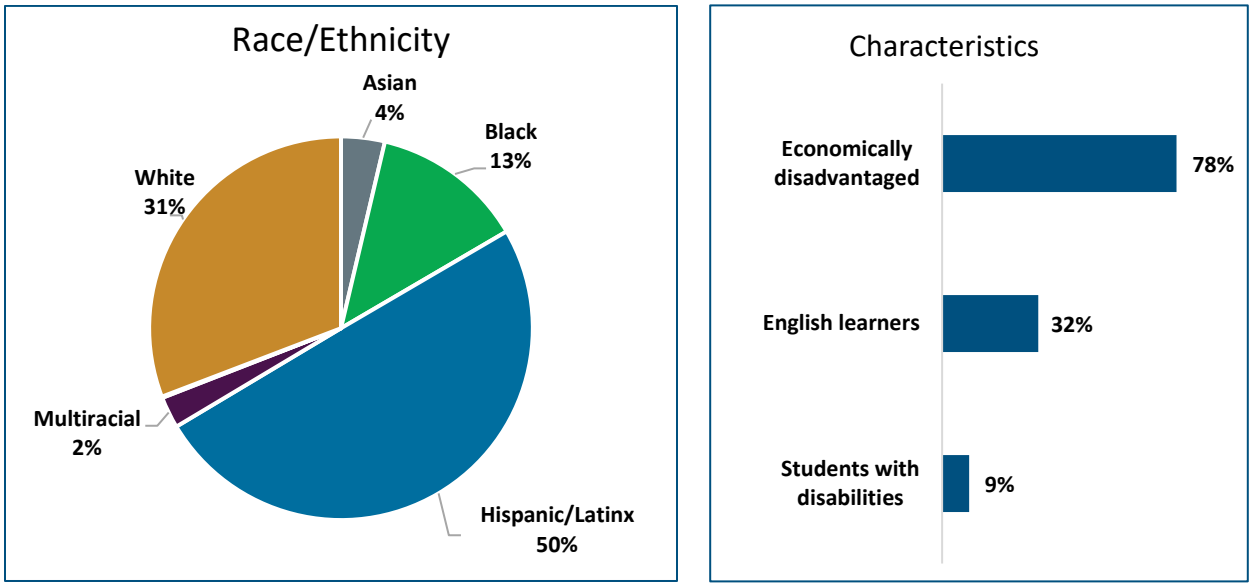
Note. N = 191. This count reflects the number of students who were matched between vendor-provided student rosters and tutoring log attendance data, and who have both pretest and posttest DIBELS composite scores.

According to the state student records, half of all students served across the three vendors were reported as Hispanic or Latinx, 31% were reported as White, 13% were reported as Black, 4% were reported as Asian, and 2% were reported as multiracial. Fewer than 1% of students served were reported as "Other."

The majority (78%) of students served across all three vendors were reported as low income, meaning either they are eligible for free or reduced-price lunch or they receive Temporary Aid for Needy Families benefits. Thirty-two percent of students served were reported as English Learners, meaning they have indicated a home language other than English, are less than proficient on an English language proficiency assessment, and are unable to perform ordinary classroom work in English. Last, 9% of students served were reported as recipients of special education services.[7] Exhibit 7 presents a snapshot of the demographic characteristics of the students served across all three vendors.

[7] See https://www.doe.mass.edu/infoservices/data/SIMS/sims-datahandbook.docx for more information.

Exhibit 7. Demographic Snapshot of the Students Included Across the Three Vendors



Note. N = 977 for both charts. The study team used a fuzzy matching program to probabilistically match the names of the students served across the merged roster, assessment, and tutoring log data, and the state student records provided by DESE. Approximately 3% of all students served could not be matched to the state student records.

Catapult

Catapult (catapultlearning.com), a national tutoring company founded in 1976, provided GEER-funded services to 438 students in 15 schools across five districts: Gardner, North Adams, Salem, Taunton, and Winchendon (Exhibit 8).

Exhibit 8. Snapshot of Catapult Tutoring Services Included in this Study

Participating Catapult Students 438	**Catapult Schools** 15	**Catapult Districts** 5
Number of Total Tutoring Sessions 8,354	**Number of Tutoring Sessions Per Student** 19	
Tutoring Attendance Per Student Per Week 2.12	**Observed Expenditure Per Student** $1,186	

Catapult uses a small-group instructional model that includes a diagnostic and prescriptive approach. Catapult tutors use formative assessments to identify students' literacy skills and needs, provide direct instruction to students, and incorporate structured lesson plans into the tutoring sessions.[8] The Catapult implementation model is flexible because it allows sites to adapt how tutoring is implemented at their school to their needs. For example, sites can choose to offer tutoring during the school day, before school, or after school. In terms of the tutors' background and qualifications, Catapult aims to hire credentialed teachers and paraprofessionals from the local community.

Catapult works closely with its partner schools, which identify and enroll students into the program. A Catapult supervisor coordinates the program for two or more districts, oversees management of the program at each participating school, and supervises all tutors at the participating districts.

DESE reported that the program provided 120 minutes of tutoring per week per student. According to program staff, tutoring sessions took place in small groups with a maximum of five students per group receiving tutoring for an estimated 40 minutes per school day. Catapult tutors who completed the survey reported that sessions occurred in groups of two to four students.[9]

Overview of Catapult Tutoring Services

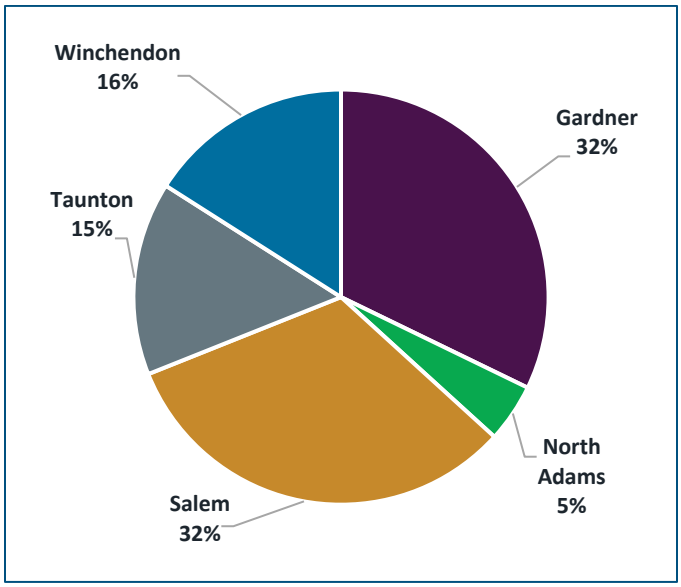
During the course of the study, Catapult tutors provided a total of 8,354 tutoring sessions. Students participated in an average of 19 tutoring sessions during the study observation window and an average of 2.12 tutoring sessions per week. The observed expenditure per Catapult student participant was $1,186.

Communities Served by Catapult

According to student roster and tutoring log data provided by Catapult, 32% of students included in the study attended Gardner Public Schools, 32% attended Salem Public Schools, 16% attended Winchendon Public Schools, 15% attended Taunton Public Schools, and 5% attended North Adams Public Schools. Exhibit 9 provides a snapshot of the home districts of the students served by Catapult.

[8] For more information about Catapult's instructional model, see https://catapultlearning.com/programs/instruction/.
[9] *N* = 48.

Exhibit 9. Home Districts of the Students Served by Catapult

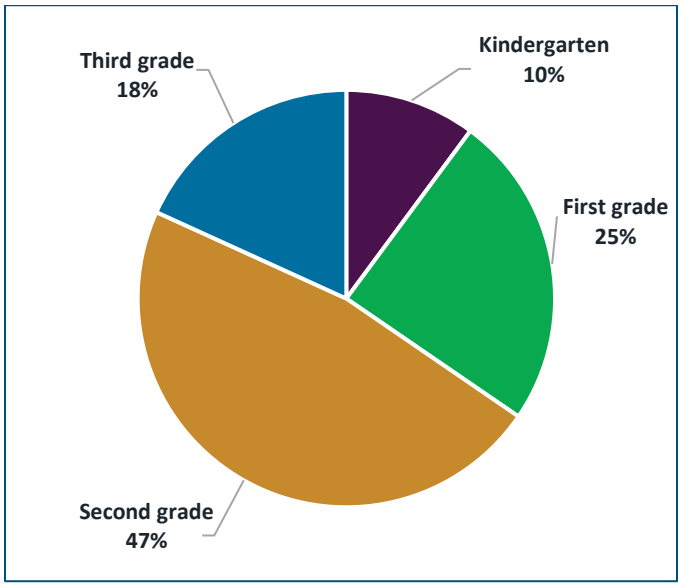


Note. N = 438.

Characteristics of the Students Served by Catapult

According to state student records provided by DESE, 10% of students served by Catapult were in kindergarten, 25% were in first grade, 47% were in second grade, and 18% were in third grade. Exhibit 10 presents a snapshot of the grade levels of the students served by Catapult.

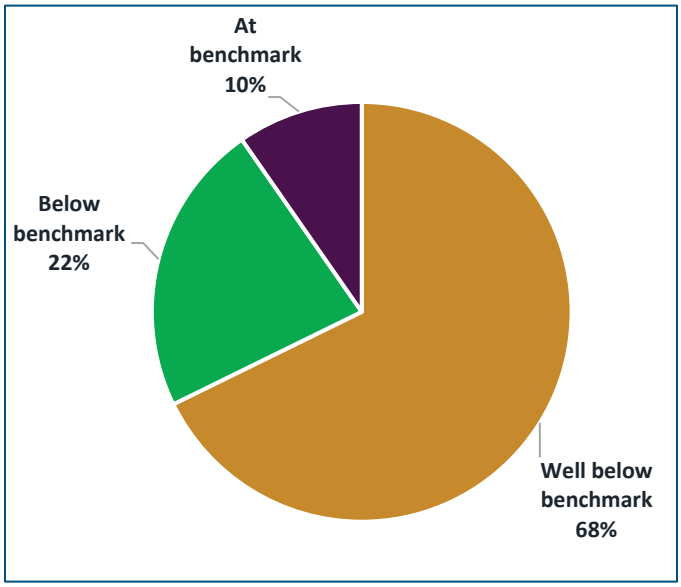
Exhibit 10. Grade Levels of the Students Served by Catapult



Note. N = 430. The study team used a fuzzy matching program to probabilistically match the names of the students served across the merged roster, assessment, and tutoring log data, and the state student records provided by DESE. Approximately 2% of Catapult students could not be matched to the state student records.

Of the Catapult students included in our analysis of learning outcomes, 68% scored well below benchmark on the baseline DIBELS assessment, 22% scored below benchmark, and 10% scored at benchmark. Exhibit 11 presents the baseline DIBELS reading levels of the students served by Catapult.

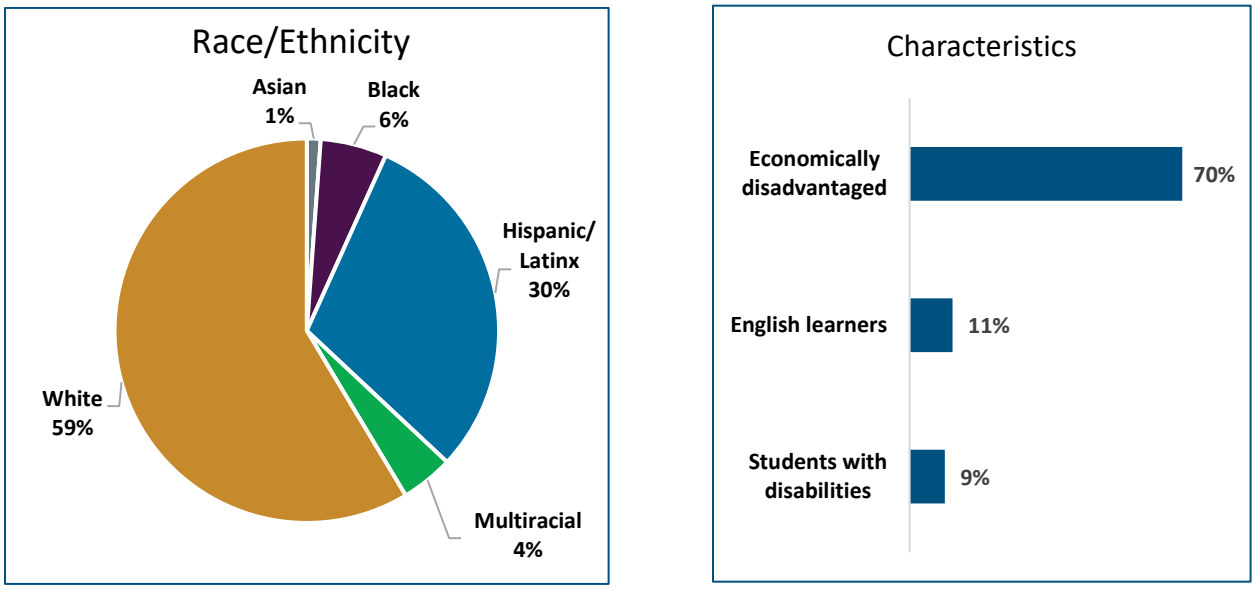
Exhibit 11. Baseline DIBELS Reading Levels of the Students Served



Note. *N* = 62. This count reflects the number of students who were matched between vendor-provided student rosters and tutoring log attendance data, and who have both pretest and posttest DIBELS composite scores.

According to the state student records, 59% of students served identified as White, 30% identified as Hispanic or Latinx, 6% identified as Black, 4% identified as multiracial, and 1% identified as Asian. Seventy percent of the students served were identified as economically disadvantaged, 11% were identified as English learners, and 9% were identified as recipients of special education services. Exhibit 12 provides a snapshot of the demographic characteristics of the students served by Catapult.

Exhibit 12. Demographic Snapshot of the Students Served by Catapult



Note. N = 430 for both charts. The study team used a fuzzy matching program to probabilistically match the names of the students served across the merged roster, assessment, and tutoring log data, and the state student records provided by DESE. Approximately 2% of Catapult students could not be matched to the state student records.

Literacy Lab

Literacy Lab (theliteracylab.org) offers individualized reading instruction to children with the goal of closing the literacy gap. Literacy Lab tutors, who are trained as part of an AmeriCorps program, focus on preventing reading failure and ensuring that all students are reading on grade level by third grade. Literacy Lab tutors work full time at elementary schools. After training, tutors provide daily one-on-one, 20-minute intervention sessions with students in kindergarten through third grade. Tutors provide targeted reading skill practice in the areas of phonemic awareness, phonics, and fluency. Tutoring sessions occur during the school day outside the student's classroom-based core reading instruction time (Exhibit 13).

Literacy Lab works closely with its partner schools, which identify and enroll students into the program. Each school provides a staff person—a literacy specialist, coach, or interventionist—to work as a program coordinator.[10] The person in this role handles communication with Literacy Lab, schedules students for the tutoring service, and serves as an on-site supervisor for the tutors at the school.

[10] Literacy Lab refers to this role as an "internal coach." To be consistent across vendors, we use the "program coordinator" title.

Tutors and the Literacy Lab leader reported that the program aimed to provide tutoring 20 minutes per day for 5 days, yielding a target dosage of 100 minutes per week per student. Literacy Lab tutors recorded all student absences from scheduled weekly sessions in the tutoring logs. According to program staff and tutor survey data, tutoring sessions took place one-on-one. Individual students received tutoring for 20 minutes per school day.[11]

Overview of Literacy Lab Tutoring Services

During the course of the study, Literacy Lab tutors provided a total of 6,545 tutoring sessions. Students participated in an average of 60 sessions during the study observation window, and an average of 2.62 tutoring sessions per week. The observed expenditure per Literacy Lab student participant was $1,909 (Exhibit 13).

Exhibit 13. Snapshot of Literacy Lab Tutoring Services Included in this Study

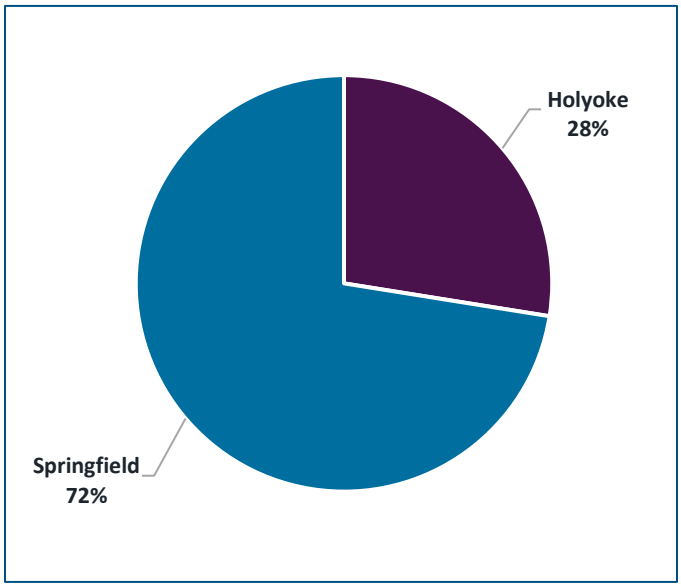
Participating Literacy Lab Students  **109**	**Literacy Lab Schools**  **7**	**Literacy Lab Districts** **2**
Number of Total Tutoring Sessions  **6,545**		**Number of Tutoring Sessions Per Student**  **60**
Tutoring Attendance Per Student Per Week  **2.62**		**Observed Expenditure Per Student**  **$1,909**

Communities Served by Literacy Lab

According to student roster and tutoring log data provided by Literacy Lab, 28% of students served attended Holyoke Public Schools and 72% attended Springfield Public Schools. Exhibit 14 provides a snapshot of the home districts of the students served by Literacy Lab.

[11] For the tutor survey data, *N* = 6.

Exhibit 14. Home Districts of the Students Served by Literacy Lab

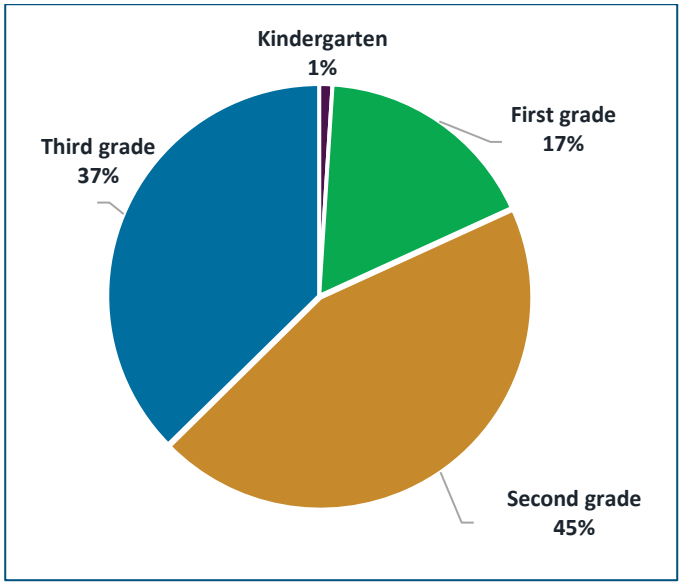


Note. N = 109.

Characteristics of the Students Served by Literacy Lab

According to state student records provided by DESE, 1% of students served by Literacy Lab were in kindergarten, 17% were in first grade, 45% were in second grade, and 37% were in third grade. Exhibit 15 presents a snapshot of the grade levels of the students served by Literacy Lab.

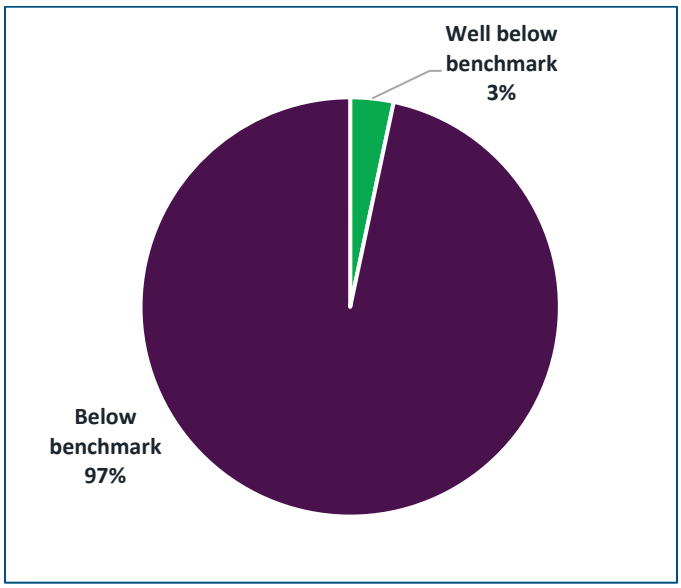
Exhibit 15. Grade Levels of the Students Served by Literacy Lab



Note. N = 99. The study team used a fuzzy matching program to probabilistically match the names of the students served across the merged roster, assessment, and tutoring log data, and the state student records provided by DESE. Approximately 1% of students served by Literacy Lab could not be matched to the state student records.

Of the Literacy Lab students included in our analysis of learning outcomes, 97% scored below benchmark on the baseline DIBELS and 3% scored well below benchmark. Exhibit 16 presents the baseline DIBELS reading levels of the students served by Literacy Lab.

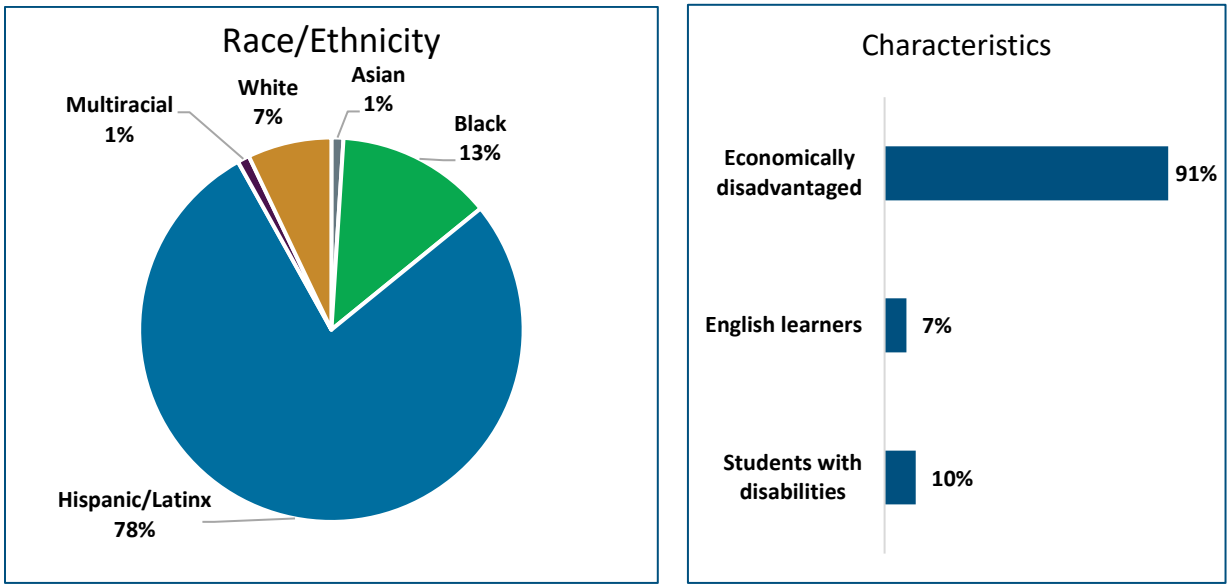
Exhibit 16. Baseline DIBELS Reading Levels of the Students Served by Literacy Lab



Note. N = 30. This count reflects the number of students who were matched between vendor-provided student rosters and tutoring log attendance data, and who have both pretest and posttest DIBELS composite scores.

According to the state student records, 78% of students served by Literacy Lab identified as Hispanic or Latinx, 13% identified as Black, 7% identified as White, and 1% each identified as Asian and multiracial, respectively. Ninety-one percent of students served were identified as economically disadvantaged, 7% were identified as English learners, and 10% were identified as recipients of special education services. Exhibit 17 provides a demographic snapshot of the students served by Literacy Lab.

Exhibit 17. Demographic Snapshot of the Students Served by Literacy Lab



Note. N = 99 for both charts. The study team used a fuzzy matching program to probabilistically match the names of the students served across the merged roster, assessment, and tutoring log data, and the state student records provided by DESE. Approximately 1% of students served by Literacy Lab could not be matched to the state student records.

Springboard

Springboard (www.springboardcollaborative.org), a national tutoring company founded in 2011, provided GEER-funded services to over 2,400 students across : Boston, Chelsea, Lawrence, Lynn, and Randolph.

Springboard's tutoring model combines a classroom curriculum with family engagement.[12] Its curriculum emphasizes phonemic awareness, phonics, fluency, vocabulary, and comprehension. Family engagement includes virtual workshops through which tutors discuss literacy skills and concepts, along with reading strategies to use at home. Tutors give students informational handouts to take home that include "reading tips."

In partnering with districts, Springboard identifies one or more program coordinators at the launch of the program. The program coordinator is a district or school employee who serves as a liaison with Springboard. The program coordinator oversees the program, monitors tutors, and helps recruit teachers and school staff to become tutors. The program coordinator is responsible for the day-to-day communication with families. Springboard hires school

[12] See https://www.springboardcollaborative.org/what-we-do/#Out_of_school_time_programs for more information.

personnel to serve in the role of tutor, giving priority to teachers, and hires other school staff, such as reading specialists, interventionists, and paraprofessionals.

Tutoring sessions occurred before or after school in groups that ranged from one to 12 students per tutor. Students participated in 90-minute tutoring sessions 2 days per week. Most (64%) Springboard tutors who completed the survey reported that sessions occurred in groups of five or more students.[13]

Overview of Springboard Tutoring Services

During the course of study, Springboard tutors provided 8,033 total tutoring sessions. Students participated in an average of 17 tutoring sessions during the study observation window and an average of 2.08 tutoring sessions per week. The observed expenditure per Springboard student participant was $925 (Exhibit 18). Note that these figures reflect only the services delivered and documented during the course of the program evaluation.

Exhibit 18. Snapshot of Springboard Tutoring Services Included in this Study

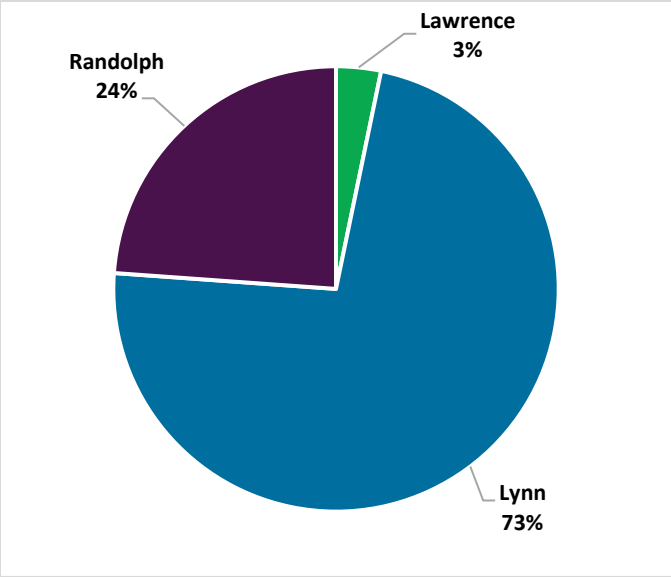
Participating Springboard Students  **461**	**Springboard Schools**  **13**	**Springboard Districts** **3**
Number of Total Tutoring Sessions  **8,033**	**Number of Tutoring Sessions Per Student**  **17**	
Tutoring Sessions Per Student Per Week  **2.08**	**Observed Expenditures Per Student**  **$925**	

Communities Served by Springboard

According to student roster and tutoring log data provided by Springboard, 73% of students included in the study attended Lynn Public Schools, 24% attended Randolph Public Schools, and 3% attended Lawrence Public Schools. Exhibit 19 provides a snapshot of the home districts of the Springboard students included in the study.

[13] *N* = 88.

Exhibit 19. Home Districts of the Students Served by Springboard

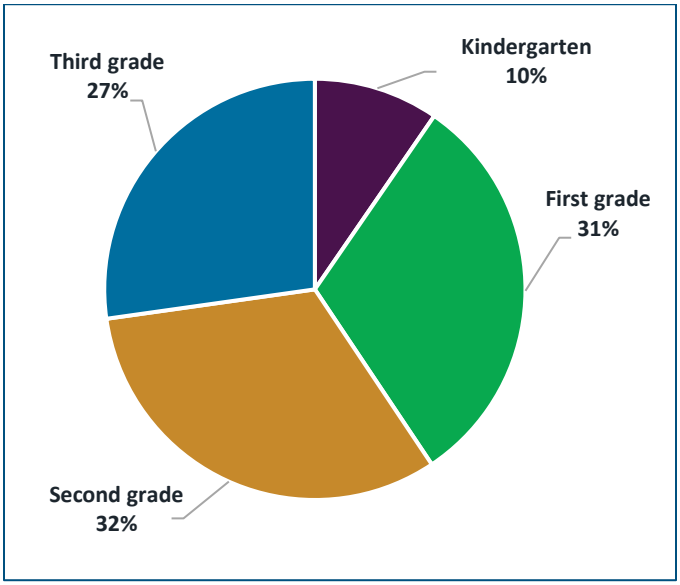


Note. N = 461.

Characteristics of the Students Served by Springboard

According to state student records provided by DESE, 10% of students served by Springboard were in kindergarten, 31% were in first grade, 32% were in second grade, and 27% were in third grade. Exhibit 20 presents a snapshot of the grade levels of the students served by Springboard.

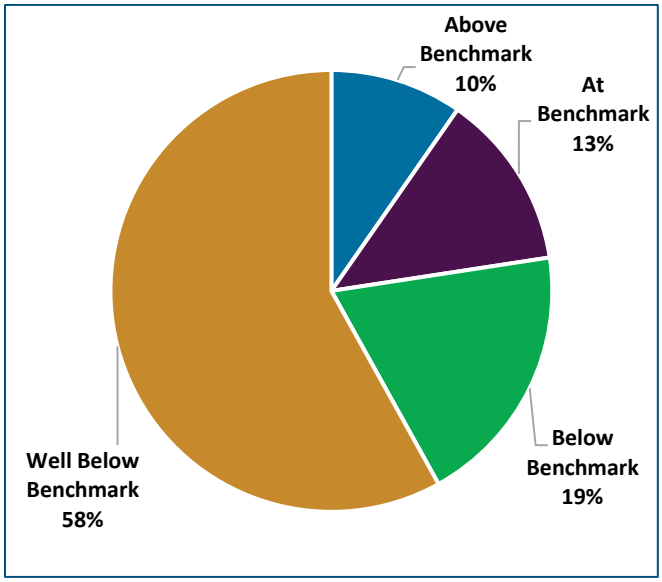
Exhibit 20. Grade Levels of the Students Served by Springboard



Note. N = 448. The study team used a fuzzy matching program to probabilistically match the names of the students served across the merged roster, assessment, and tutoring log data, and the state student records provided by DESE. Approximately 3% of students served by Springboard could not be matched to the state student records.

Of the Springboard students included in our analysis of learning outcomes, 58% scored well below benchmark on the baseline DIBELS assessment, 19% scored below benchmark, 13% scored at benchmark, and 10% scored above benchmark. Exhibit 21 presents the baseline DIBELS reading levels of students served by Springboard.

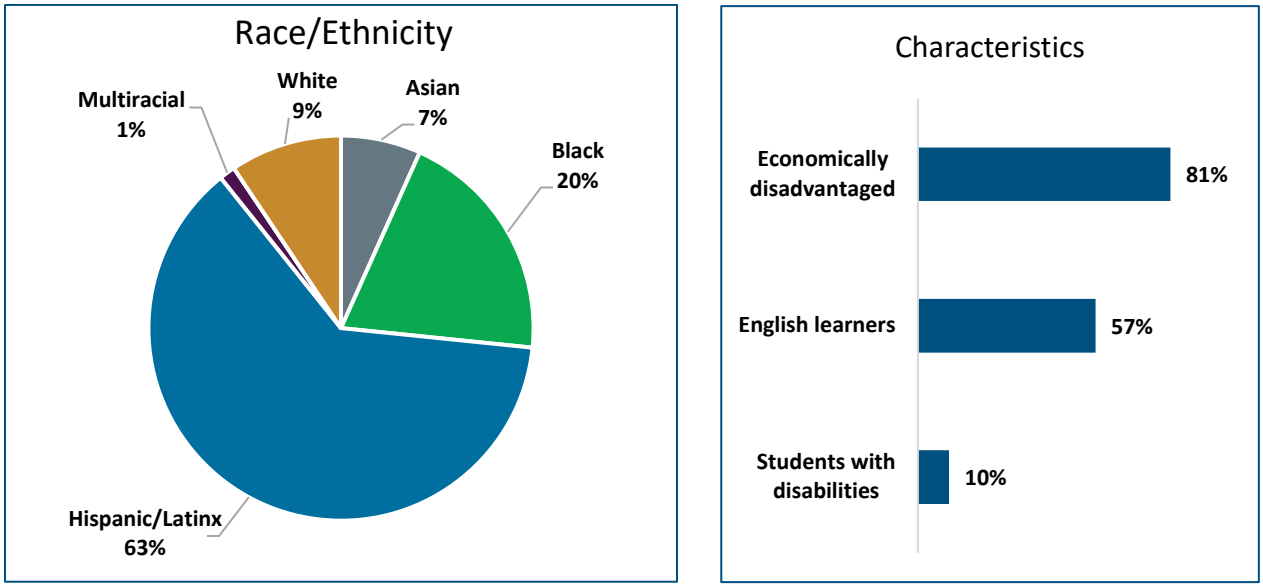
Exhibit 21. Baseline DIBELS Reading Scores of the Students Served



Note. N = 99. This count reflects the number of students who were matched between vendor-provided student rosters and tutoring log attendance data, and who have both pretest and posttest DIBELS composite scores.

According to the state student records, 63% of students served by Springboard identified as Hispanic or Latinx, 20% identified as Black, 9% identified as White, 7% identified as Asian, and 4% identified as multiracial. Fewer than 1% identified as "Other." Additionally, 81% of students served by Springboard were identified as economically disadvantaged, 57% were identified as English learners, and 10% were identified as recipients of special education services. Exhibit 22 provides a snapshot of the demographic characteristics of the students served by Springboard.

Exhibit 22. Demographic Snapshot of the Students Served by Springboard



Note. N = 448 for both charts. The study team used a fuzzy matching program to probabilistically match the names of the students served across the merged roster, assessment, and tutoring log data, and the state student records provided by DESE. Approximately 3% of students served by Springboard could not be matched to the state student records.

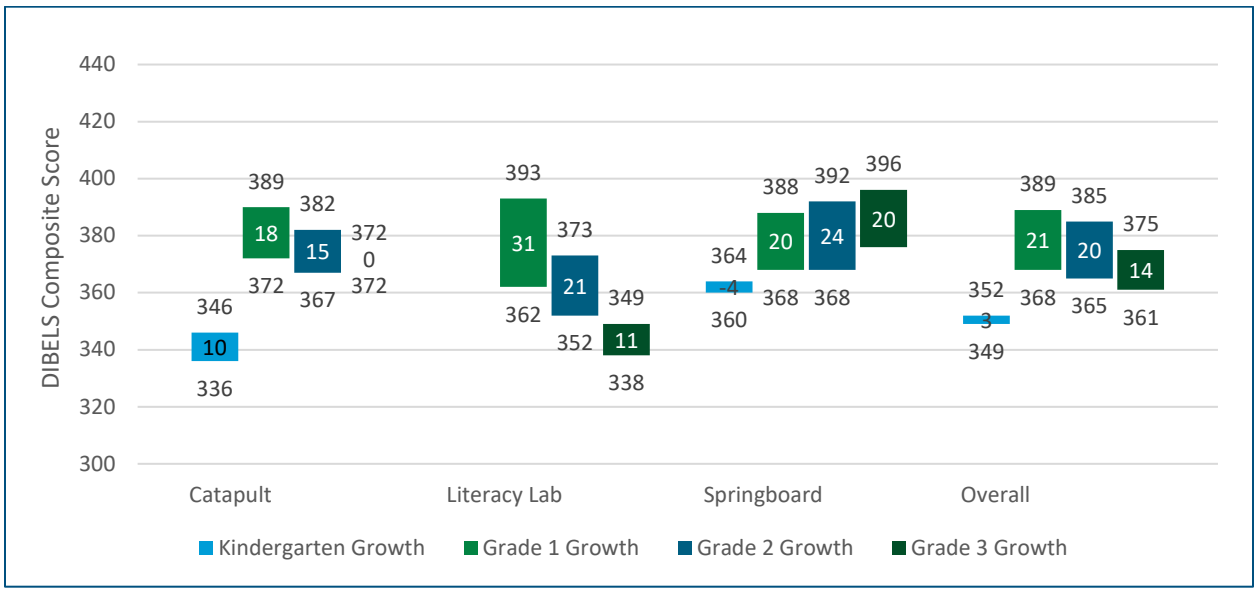
Summary of Student Learning Outcomes

In this section, we summarize student learning outcomes. We begin with descriptive analyses, and then proceed to regression analyses that test for associations across program features, student characteristics, and learning outcomes. On average, students' composite scores on the DIBELS increased by 16 points (standard deviation [*SD*] = 18). Catapult students' scores increased, on average, 13 points (*SD* = 15); Literacy Lab students' scores increased 17 points (*SD* = 17); and Springboard students' scores increased 17 points (*SD* = 19).

Exhibit 23 presents the average DIBELS pretest scores, posttest scores, and score growth, for each combination of tutoring vendor and grade level. The numbers below the bars are the average pretest scores, the numbers in the bars are the average growth, and the numbers above the bars are the average posttest scores. The clusters of bars correspond to vendors, and the colors of the bars correspond to grade levels. For instance, the left-most light blue bar presents results for kindergartners at Catapult tutoring sites; the average pretest score among kindergartners at Catapult sites is 336 points, and the average posttest score among kindergartners at Catapult sites is 346 points, corresponding to an average growth of 10 points. Note that there were not enough kindergartners at Literacy Lab sites with complete data to

present results for that grade-vendor combination. A student reading at their expected grade level would earn a score around 400 at the midpoint of the school year; the nationwide standard deviation of test scores is 40 points at any given point in time.

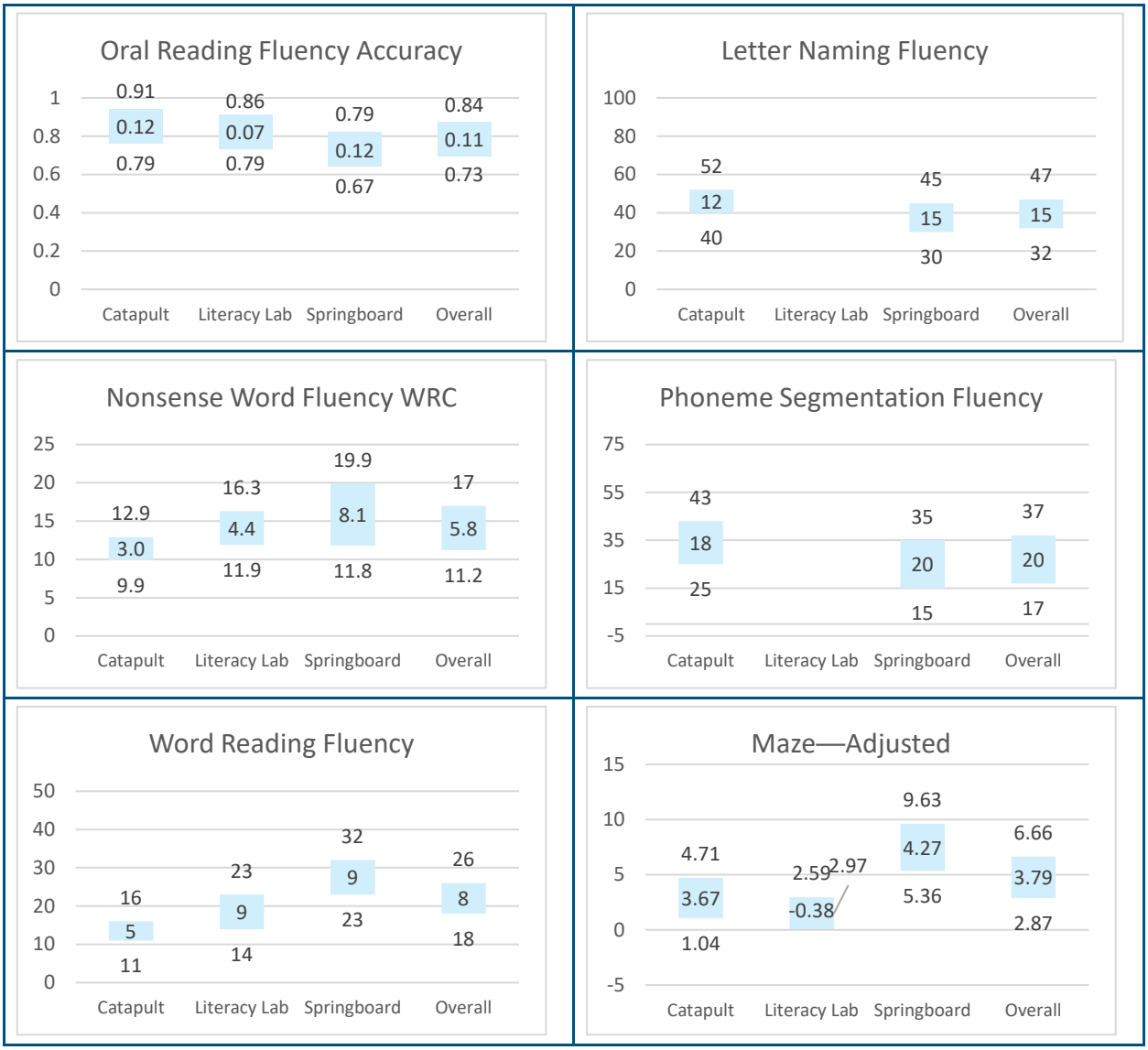
Exhibit 23. Average Student Growth on DIBELS Composite Scores



Approximately 67% of students with complete data had baseline DIBELS scores well below benchmark. At posttest, 45% scored well below benchmark. At baseline, 15% of students had DIBELS composite scores at or above benchmark; this increased to 37% at posttest.

Exhibit 24 presents students' average pretest scores, posttest scores, and score growth for each of the six DIBELS subscales, separately by vendor. Springboard students made particular gains in Nonsense Word Fluency and the Maze. These subscales measure the alphabetic principle and comprehension, respectively. In general, however, students' growth on each subscale was largely similar across vendors. Letter Naming Fluency and Phoneme Segmentation Fluency scores are not shown for Literacy Lab due to sample size constraints.

Exhibit 24. Average Student Growth on DIBELS Scores, by Subscale



Associations Between Program Characteristics and Student Outcomes

RQ1. Was there a significant difference between the impact of providers' program models on student outcomes?

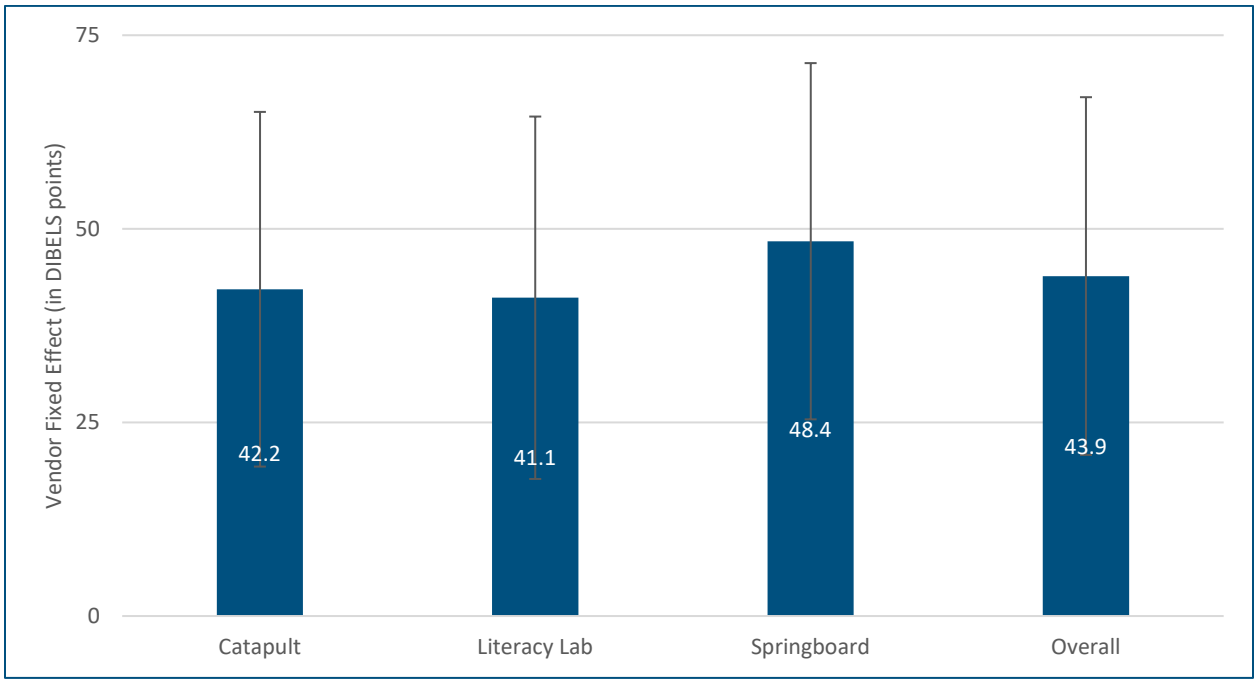
All vendors' tutoring models were equally successful in supporting student learning (see Exhibit 25). Fixed effects regression models were utilized to examine student growth from DIBELS pretest to posttest. Fixed effects regression models are a measure of the impact of the program after controlling for other factors. These models indicate that students gained between 41 and 48 points in each tutoring program. These statistical models included controls for students' demographics and *excluded* specific details of service delivery (e.g., number of tutoring

sessions, whether the tutor was a teacher, etc.). This approach allowed us to isolate the association between student learning and each vendor's program model as a whole package.

While this is a common analytic approach to compare programs, it is important to keep in mind that these estimates assume values of 0 on all covariates, and a score of 0 on the DIBELS is impossible. The lowest observed DIBELS pretest score is 243. This modeling technique allows us to compare the three vendors with each other. Doing so, we see that the three vendor-specific estimated effects are remarkably similar to each other. No differences between any pair of vendor effects are statistically significant at any conventional threshold.

These results are notable. We might have expected Literacy Lab students to grow more than the students working with Catapult or Springboard because they received more tutoring between pretest and posttest, on average. Thus, Literacy Lab students may have had more time to learn skills, both in class and during tutoring sessions. However, again, the results indicate that student learning was equivalent across the three programs.

Exhibit 25. Vendor-Specific Effects on Student Outcomes



RQ2. Which program factors best predicted student growth?

Exhibit 26 contains the relationships between various aspects of program delivery and student DIBELS composite score growth. Students who attended more hours of tutoring had larger score gains; for instance, a student who attended 20 hours of tutoring would grow by about 4 more points than a student who attended 10 hours of tutoring. Other variables with significant

relationships with DIBELS score growth were an indicator that a student's tutor is a certified teacher and an indicator that a student's tutor received frequent feedback; perhaps surprisingly, both of these were associated with lower score growth. However, because the three vendors had program models that were delivered consistently but were very different from each other, the values of the program variables largely signify the vendor that administered a student's tutoring. For instance, the vast majority of certified teachers involved in tutoring worked with Springboard, and none worked with Literacy Lab, which hired AmeriCorps members. Springboard held its tutoring after school, whereas Literacy Lab held it during school hours. Therefore, the "effect" of working with Springboard can be approximated through the combination of the effects of having a certified teacher and holding tutoring after school, but it is difficult to separate the effects of those two program characteristics from each other.

Exhibit 26. Relationships Between Program Factors and Student Growth

Program factor	Coefficient (standard error)
Number of students in tutoring group	0.91 (3.34)
Tutor is a certified teacher	–16.09 (6.34)*
Number of sessions per week	5.30 (3.04)
Number of hours of training tutor received	0.01 (0.57)
Tutor received feedback often	–11.52 (4.12)*
Tutoring takes place outside of school hours	7.58 (5.33)
Number of hours of tutoring received	0.42 (0.20)*

*$p < .05$.

RQ3. Were there significant differences in how students developed specific skills that were impacted by program factors?

To understand the potential differences in how students in each program developed specific skills, we conducted regression analyses that used each of the DIBELS subscale scores as outcomes. Most program factors were not significantly related to any of the subscale scores. A consistent pattern did not emerge among the few program factors that did predict growth on the DIBELS subscale scores. Students whose tutors frequently received feedback had lower growth on the Nonsense Word Fluency, words read correctly (WRC) subscale, whereas students who received more hours of tutoring showed more growth on the Word Reading Fluency section and in Maze accuracy. Students whose tutors were teachers had significantly lower growth in the number of correct answers in Oral Reading Fluency and in Maze accuracy. Students who attended more sessions per week had a larger reduction in the number of errors

on the Oral Reading Fluency section but lower growth in the Maze section in terms of both the number and proportion of correct answers. As with the associations between program factors and overall DIBELS score growth, the interpretation of any of these results, except perhaps for the number of hours, is complicated by how much the program factors are distinctive aspects of particular vendors, making it difficult to distinguish the effect of having a tutor who is a teacher from the effect of being enrolled with Springboard. The results involving the number of hours of tutoring suggest that the benefits of practice time may be greatest in Word Reading Fluency and in the Maze test.

Associations Between Student Characteristics and Student Outcomes

RQ4. Which student characteristics best predicted student growth?

Three student characteristics have statistically significant relationships with student growth: grade level, race/ethnicity, and pretest DIBELS score (see Exhibit 27). Kindergartners had lower DIBELS score growth than students in other grades, although the scoring range for kindergarten is also more condensed than in other grades. Hispanic or Latinx students improved their DIBELS scores more than otherwise-identical White or Black students exposed to equivalent tutoring services, although not significantly more than Asian students. Additionally, students with lower pretest scores grew more on the DIBELS than students with higher baseline scores. While these students have more room to grow than higher-scoring students, who have fewer points left to gain, few students score near the theoretical maximum DIBELS composite score. Thus, it is unlikely that proximity to the ceiling is the primary factor that explains the negative relationship between pretest scores and score growth. This result suggests evidence that the tutoring program is succeeding in its goal of narrowing test score gaps and bringing lower-scoring students closer to grade level.

Exhibit 27. Relationships Between Student Characteristics and Student Growth

Program factor	Coefficient (standard error)
Student is in kindergarten	–26.34 (5.95)*
Student is in first grade	0.00 – reference category
Student is in second grade	–2.63 (4.92)
Student is in third grade	–6.86 (7.43)
Student is Black	–1.60 (3.16)
Student is Hispanic/Latinx	8.46 (3.31)*
Student is Asian	9.72 (5.20)
Student is an English learner	2.75 (3.47)
Student is eligible for free or reduced-price lunch	–3.98 (2.93)

Program factor	Coefficient (standard error)
Student is in special education	0.58 (3.67)
DIBELS pretest score	–0.11 (0.05)*

*$p < .05$.

RQ5. Were there significant differences in how students developed specific skills that were impacted by student characteristics?

The regression specifications with DIBELS subscale scores as outcomes included student characteristics, in addition to program characteristics, as covariates. As in the specifications with DIBELS composite scales as outcomes, students with lower baseline subscale scores had higher score growth in most subscales, including Letter Naming Fluency, Phoneme Segmentation Fluency, Oral Reading Fluency accuracy, and Maze correct answers. Economically disadvantaged students had lower growth in Phoneme Segmentation Fluency, Word Reading Fluency, and Nonsense Word Fluency.[14] Asian students had lower growth in Phoneme Segmentation Fluency, and Hispanic or Latinx students had higher growth in Word Reading Fluency but lower growth in Oral Reading Fluency accuracy. Kindergartners had lower growth in Nonsense Word Fluency, and second graders had higher growth in Word Reading Fluency.

What Works for Whom, and Under Which Conditions?

RQ6. What significant interaction effects were found between any of the variables in RQs 1–5?

We tested interactions that combined DIBELS pretest scores with indicators for each of the three vendors. If the coefficient on this interaction term is negative and significant for a particular vendor, it implies that the vendor is making progress in helping lower-achieving students catch up to grade level. However, a coefficient is positive and significant implies that the vendor is helping higher-achieving students grow faster while students with less-developed reading backgrounds fall further behind their peers. This interaction term is statistically significant for all three vendors. For Catapult and Springboard, the coefficient is negative, suggesting the gaps were closing. For Literacy Lab, the coefficient is positive. The negative coefficient suggests the disparities were widening, but additional analyses would be needed to confirm this finding.

[14] Economically disadvantaged students scored similarly on the pretest to students who were not economically disadvantaged, meaning that higher growth among students with low pretest scores can be interpreted separately from lower growth among economically disadvantaged students rather than assuming the groups are the same and the opposing effects of being in each group cancel out.

To determine whether different program characteristics were more or less helpful for particular types of students, we reestimated the main model separately with each of the following interaction terms added one by one:

- The number of hours of training received by the tutor, interacted with whether the tutor is a teacher or another type of educator.
- The size of groups in which the student works, interacted with whether the tutor is a teacher or another type of educator.
- The size of groups in which the student works, interacted with indicators for each grade level.
- Whether the tutor and student are both Black or both Hispanic or Latinx. (There were not sufficient data to support the inclusion of an interaction term for the tutor and student both being Asian.)

While there are theoretical justifications for each of these interactions being relevant to student growth, only one of these interactions had a statistically significant relationship with DIBELS score growth: Students in kindergarten who worked in smaller groups had higher DIBELS score growth than kindergartners who worked in larger groups. However, the direction of causation in this case is not immediately obvious.

Cost Analysis

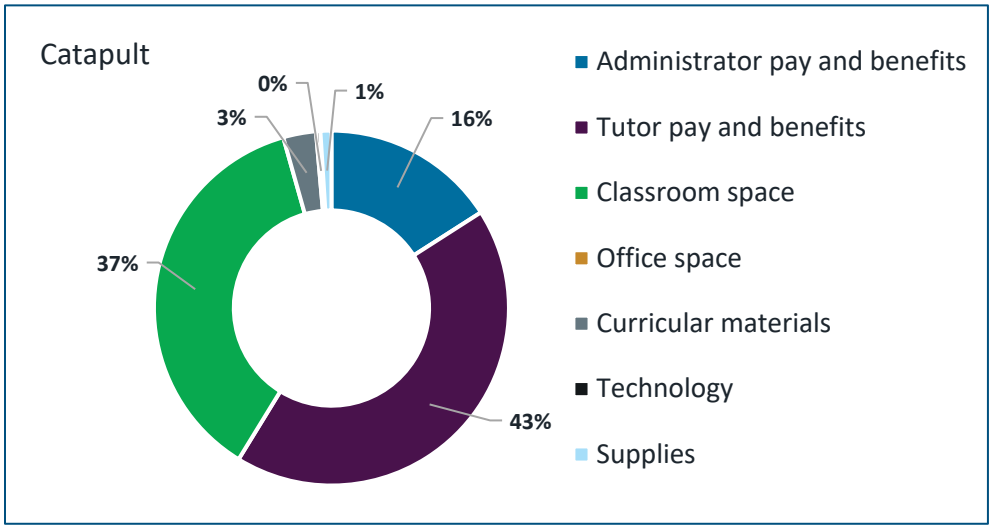
RQ7. What is the most cost-effective approach to future implementation and funding of early literacy programming?

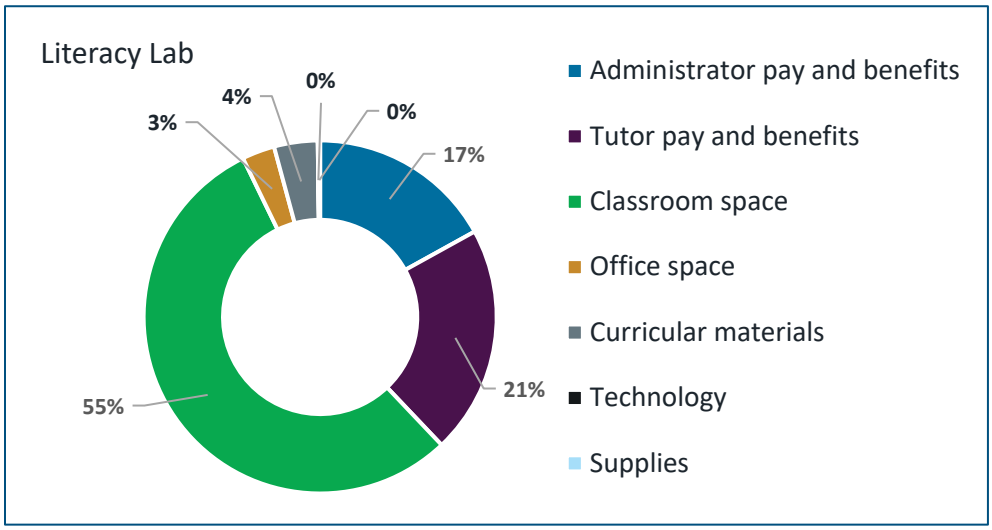
While students working with all three tutoring vendors had substantial growth in their DIBELS scores during the course of the year, the three vendors implemented very different models, which is noticeable in the breakdown of the expenses that would be required to replicate them in the future (see Exhibit 28). The expenses required to replicate programs are often not the same as the costs of implementing them during the period of study[15], as some resources may have been donated or purchased at a discount; however, there is no guarantee that these options would be available in all contexts in which the programs could be implemented in the future, so these calculations assume that full market value (or a market value scaled by the proportion of time used, in the case of shared resources such as school-owned computers) would be paid for all inputs in the future. Catapult and Springboard relied more on certified teachers, who would be more expensive if paid the hourly equivalent of their teaching salaries, whereas Literacy Lab used AmeriCorps members who received lower pay. In addition, Catapult

[15] The costs of implementing each program during the period of study are referred to as "observed per-student expenditure in 2021-22" in Exhibit 29.

asked tutors to work more hours per week and more weeks than Literacy Lab and Springboard. Meanwhile, Literacy Lab used dedicated classroom space, which composed a large fraction of its expenses. Literacy Lab was the only vendor that used any office space. Technology and supplies composed less than 3% of any vendor's replication expenses. Considering that DIBELS score growth was similar across vendors, Springboard's lower per-student expense to replicate makes it appear to be by far the most cost-effective tutoring model to expand in the future (see Exhibit 29).

Exhibit 28. Estimated Replication Cost Breakdowns, by Vendor





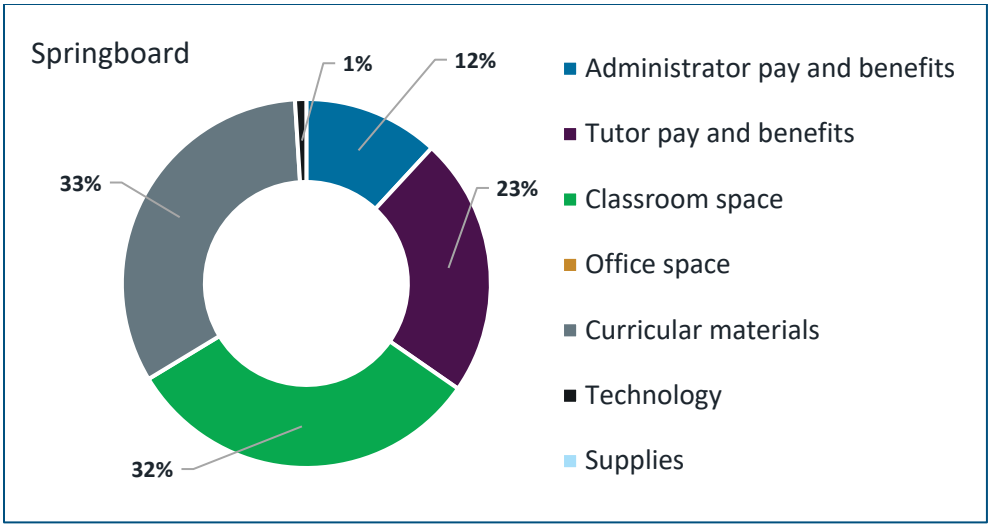


Exhibit 29. Summary of Replication Cost and Cost-Effectiveness Findings

	Catapult	Literacy Lab	Springboard
Number of students served during the study	700	135	637
Estimated total cost to replicate program exactly	$4,059,204	$615,974	$1,759,709
Estimated per-student cost to replicate program exactly	$5,799	$4,563	$2,762
Per-student expenditure in 2021-22	$1,186	$1,909	$925
Estimated cost to replicate program for 200 total students	$1,135,303	$912,554	$505,262
Estimated cost to improve one future student's performance by 0.25 *SD* (10 DIBELS points)	$4,617	$2,743	$1,583

Limitations

Several limitations of this study should be kept in mind when interpreting its findings. As noted, many program characteristics were very specific to individual vendors. Variation within vendors was minimal but variation across vendors was substantial. As such, program characteristics were often highly correlated with each other, making it very difficult to distinguish the effects of different aspects of the vendors' program delivery. In addition, without a comparison group that receives only regular classroom instruction, we cannot determine how much student learning is due to the tutoring services versus regular classroom instruction.

Additionally, data were available for only a subset of the students served by the three tutoring vendors. The analytic sample includes only students with DIBELS pretest and posttest scores who can be matched to vendors' student rosters, attendance logs, and state administrative data. This leaves out students in schools that did not administer DIBELS in a timely fashion (or

at all), students whose tutors did not record attendance data, and students who were absent when the DIBELS was administered. Because several large districts are excluded from the analytic sample as a result, such as Boston (which did not administer DIBELS) and Lynn (which administered the Renaissance STAR instead), the analytic sample is smaller than the population and not entirely representative of the population. This limits the statistical power of the analyses and the extent to which findings can be extrapolated to other settings where tutoring took place.

Furthermore, the estimated costs do not always represent the amount of money actually spent by the tutoring vendors, or that would necessarily be spent if a new organization tried to start an equivalent tutoring program. For instance, the cost calculations include the price to rent space equivalent to the amount of classroom or library space used for tutoring, but it is unlikely that tutoring vendors had to pay for space by the square foot. Instead, the provision of space may be negotiated by vendors and districts as part of a broader agreement, or provided as an in-kind contribution by districts. Similarly, resources such as books, art supplies, technology may be donated rather than purchased. Salaried teachers are unlikely to be paid the exact hourly equivalent of their teaching salaries by the tutoring companies. Thus, the cost estimates should be interpreted as the market value of all of the ingredients that it would take to reproduce each vendor's model, rather than the literal cost that each vendor paid.

Conclusion

This summative evaluation aimed to identify the program and student characteristics among students who participated in the Early Literacy Supplemental Services Tutoring program in addition to analyzing the program cost.

- **Across all tutoring vendors, students made gains in literacy learning that exceeded expected learning rates.** All tutoring vendors were equally successful in supporting students' literacy learning, according to statistical models. It was difficult to tease apart the value of specific program components or approaches because all three models of tutoring studied are distinct, and they are implemented in packages.
- **Tutoring contributed to the narrowing of test score gaps.** Students with lower baseline scores improved their test scores more than students with higher baseline scores did while enrolled in tutoring, supporting the program's goal of helping struggling readers catch up to grade level.
- **The program had biggest impact on the students it was designed to serve.** Students with lower initial DIBELS scores made the largest gains during the evaluation period.

References

Massachusetts Department of Elementary and Secondary Education (DESE). (2019). *Next generation MCAS achievement results*. https://profiles.doe.mass.edu/statereport/nextgenmcas.aspx

Nickow, A., Oreopoulos, P., & Quan, V. (2020). *The impressive effects of tutoring on preK–12 learning: A systematic review and meta-analysis of the experimental evidence*. National Bureau of Economic Research. https://www.nber.org/system/files/working_papers/w27476/w27476.pdf

Appendix A. Variable Sources and Measurement Properties

Variable	Source	Measurement properties
Number of hours of tutoring	Tutoring logs	Continuous: The number of sessions attended multiplied by the length of the session, in hours (e.g. a student attending three 45-minute sessions would be coded as 2.25)
Frequency of tutoring	Tutoring logs	Continuous: The number of sessions divided by the number of weeks between the first and last sessions
Student-tutor race match	Tutor survey, student administrative data (ideally student rosters but it is often missing)	Binary: 1 if the student and the tutor are the same race/ethnicity, 0 otherwise
Total training hours received	Tutor survey	Integer: The two questions about tutor training (summer and school year) would be coded to numbers of hours and summed (e.g., if a tutor received "more than 4, up to 8 hours" of training in the summer and "more than 2, up to 4" in the school year, they would be coded as 5 + 3 = 8)
Tutor is a certified teacher	Tutor survey	Binary: 1 if the tutor is a certified teacher and 0 otherwise
Tutor is an educator but not a teacher	Tutor survey	Binary: 1 if the tutor lists any district employment besides teaching, 0 if the tutor is a teacher or not employed by a district
Feedback received	Tutor survey	Integer: 3 if the tutor receives feedback every session, 2 if the tutor receives feedback often, 1 if the tutor receives feedback sometimes, 0 if the tutor never receives feedback or is not observed
School fixed effects	Student rosters	Binary: Indicator for each elementary school served
Vendor fixed effects	Student rosters	Binary: Indicator for each tutoring vendor
Student demographics; economic disadvantage, English learner status, special education status, age	Administrative data	Binary, except age, which would be continuous
DIBELS pretest score	DIBELS data	Continuous
Tutoring during or after the school day	Interviews/document review	Binary: 1 if after school, 0 if during school
Tutor/student ratio	Tutoring logs; survey	Continuous: number of students served by student's tutor Binary: 1 for group, 0 for individual
Student grade level	Student rosters	Categorical: Separate indicators for each value
Differentiation of tutoring activities	Tutor survey	Binary: 1 if the activities are differentiated, 0 otherwise

About the American Institutes for Research

Established in 1946, with headquarters in Arlington, Virginia, the American Institutes for Research® (AIR®) is a nonpartisan, not-for-profit organization that conducts behavioral and social science research and delivers technical assistance to solve some of the most urgent challenges in the U.S. and around the world. We advance evidence in the areas of education, health, the workforce, human services, and international development to create a better, more equitable world. For more information, visit AIR.ORG.

AIR® Headquarters
1400 Crystal Drive, 10th Floor
Arlington, VA 22202-3289
+1.202.403.5000 | **AIR.ORG**