# SIF EXPERIENCE CORPS NATIONAL EVALUATION

## Evaluation of Sustained Small Group Tutoring

**DECEMBER 2019 | 2018-2019**


SUBMITTED TO
AARP Foundation Experience Corps
601 E Street NW
Washington, DC 20004

SUBMITTED BY
Allan Porowski, Kerry Hofer,
Michele de Mars, David Rodriguez, &
Michelle Thompson, Abt Associates

Catherine Darrow,
Education Development Center

6130 Executive Blvd
Rockville, MD 20852


Abt ASSOCIATES

AARP Foundation®
EXPERIENCE CORPS
Guiding the next generation of readers

## About This Report

This report summarizes findings from an evaluation of the AARP Foundation's Experience Corps sustained small group tutoring strategy. This evaluation was funded by the Corporation for National and Community Service's Social Innovation Fund and supported by the AARP Foundation.

The authors would like to acknowledge and thank Kathy Brennan, Susan Siegel, and David Handy from the AARP Foundation for their continued guidance, and Experience Corps staff at all SIF-funded local programs who made this project study possible. We would also like to acknowledge the contributions of Sandra Wilson who provided technical guidance on this work.

## Authors

Allan Porowski, Abt Associates

Kerry Hofer, Abt Associates

Michele de Mars, Abt Associates

David Rodriguez, Abt Associates

Michelle Thompson, Abt Associates

Catherine Darrow, Education Development Center


Abt Associates | 6130 Executive Boulevard | Rockville, MD 20852

# Contents

## Exhibits

# EXECUTIVE SUMMARY

This evaluation, funded by the Corporation for National and Community Service's Social Innovation Fund, is designed to determine to what extent the ***Experience Corps sustained small group tutoring*** approach improves young children's literacy achievement.

Abt Associates conducted a randomized controlled trial during the 2018-19 school year in three AARP Foundation Experience Corps local programs: (1) Generations Inc. (Boston, MA); (2) Read to Succeed Buffalo (Buffalo, NY); and (3) United Way Central Georgia (Bibb County/Macon, GA). Randomization was conducted at the classroom level; within randomly assigned classrooms, eligible students who consented to be in the impact study were included in the analytic sample.

Altogether, 850 students in kindergarten through grade 3 from 238 classrooms in 21 schools from four school districts were included in the analytic sample. Students in the study were assigned either to a treatment group (receiving Experience Corps sustained small group tutoring) or to a control group (not receiving Experience Corps sustained small group tutoring).

Students were administered subtests of Lexia's RAPID™ literacy assessment at the beginning and end of the study period. Altogether, five subtests were administered in both the fall and spring (Phonological Awareness, Letter Sounds, Word Reading, Vocabulary Pairs, Following Directions, and Spelling), and each student took between two and four of these subtests, depending on their grade level. Average scores of the treatment and control groups were compared. Social-emotional learning was also measured using an eight-item version of the Devereux Student Strengths Assessment (the DESSA-mini). Findings were analyzed to explore four key questions of interest:

The AARP Foundation Experience Corps program aims to help K-3 students in high-need communities improve their literacy skills by matching senior volunteers in the community (age 50 or older) with students who are struggling to read.

**Conclusion:**

Although the evaluation did not uncover a statistically significant positive effect of sustained small group tutoring on literacy or social emotional skills, there is evidence in the findings of a pattern of growth in literacy skills that slightly favors the Experience Corps group. Dosage was found to be a significant predictor of positive outcomes. Fidelity to the program model varied across the three local programs, and the local program with the highest level of fidelity to the Experience Corps model showed the strongest findings.

### Question 1: What does the Experience Corps sustained small group tutoring model look like?

On average, students in the treatment group received 34.2 sessions, with 15.8 total hours of sustained small group tutoring. Although these figures align closely to AARP Foundation dosage recommendations (i.e., at least 35 sessions per year, with a minimum of 26 minutes per session), there was substantial variation in the delivery of Experience Corps sustained small group tutoring across local programs.

The average small group size ranged from 2.7 students in Macon, to 3.2 students in Buffalo, to 3.5 students in Boston. The use of the recommended Experience Corps session structure also differed

substantially at the local program level. In Macon, the local program adhered more closely to the recommended session structure than the local programs in Boston or Buffalo.

The AARP Foundation Experience Corps model evolved over the term of this evaluation. One key change to the model that took place in the summer of 2018 was the focused attention on fluency skills practice under the close monitoring and guidance from the tutor. The programs participating in this evaluation were encouraged to continue implementing sustained tutoring without conforming to the new model. Since only one of the three local programs conformed to the new model, this study does not provide a full test of the refined AARP Experience Corps model.

**Question 2: How did students who received Experience Corps sustained small group tutoring compare to the control group on literacy achievement and social-emotional learning outcomes?**

Students who received Experience Corps sustained small group tutoring did not have significantly higher literacy scores than the control group by the end of the school year, controlling for pretest scores and demographics. Students who received tutoring began the school year with substantially lower achievement levels than the control group; however, they experienced upward trajectories in reading achievement scores comparable to the control group students' trajectories. Exploratory analyses revealed a statistically significant relationship between dosage and outcomes: students who spent more time in sustained small group tutoring had significantly stronger outcomes on the RAPID. No statistically significant differences were found on social-emotional learning outcomes between students who received Experience Corps sustained small group tutoring and those who did not.

Although statistically significant findings were not observed on the RAPID or the DESSA-mini, results from a survey of participating teachers suggested that Experience Corps was effective. A large majority of teachers in all three local programs reported improvements in students' literacy skills, improvements in students' social-emotional learning skills, and that students received a sufficient number of tutoring sessions.

**Question 3: Under what conditions is Experience Corps sustained small group tutoring most effective?**

Exploratory analyses that pooled all RAPID subtest results revealed that overall, Experience Corps sustained small group tutoring was significantly more effective in improving literacy achievement in Macon than in Boston. Differences in literacy achievement between Macon and Buffalo were not statistically significant. These differences may be explained by differences in program fidelity or by differences between experimental and control groups in baseline literacy levels.

In Macon, students in treatment and control groups scored similarly at baseline on the RAPID's benchmark predictor of reading achievement—the Reading Success Probability (RSP) score—which reflects the probability that students will read at grade level by the end of the school year. In Buffalo and Boston, students' average RSP scores at baseline were lower for the treatment group than the control group, suggesting that students in Experience Corps had a lower probability of reading at grade level than the control group *before the study began*. Baseline differences between treatment and control groups could have masked differences in outcomes due to the sustained small group tutoring.

There was no systematic pattern of gender, English learner status, grade level, race/ethnicity, or special education status being significantly related to RAPID performance; and the presence of additional formal literacy support programs in classrooms did not change the overall impact results.

The findings from the experimental study do not align with past research, which showed positive effects of Experience Corps on literacy achievement using a one-on-one tutoring strategy (Lee, Morrow-Howell,

Jonson-Reid, & McCrary, 2010). Despite this disjuncture of current study findings with teacher perceptions and past research findings, the positive relationship between dosage and outcomes, coupled with the positive findings from Macon, suggest there may be a systematic relationship between the Experience Corps model delivery and outcomes. More research is needed to unpack these findings.

# Background

Reading at grade level by the end of grade 3 has been linked through research to success in school, productive work, healthy life choices, and reduced incarceration. Grade 4 students who can't read at grade level are four times less likely to graduate from high school. In 2009, more than 80% of students from low-income families (and more than 60% of students nationwide) did not hit that grade-level mark by grade 4, putting them at risk for dropping out of high school and leaving the door open for continued challenges (The Annie E. Casey Foundation, 2010).

The AARP Foundation Experience Corps program aims to address this challenge by matching senior volunteers in the community (age 50 or older) with students who are struggling to read in grade K-3 classrooms. Volunteers provide literacy tutoring through two primary strategies: (1) sustained tutoring in a one-on-one or small group setting, and (2) whole-class literacy assistance. Prior randomized controlled trials of Experience Corps found that one-on-one tutoring resulted in improved passage comprehension and general reading skills (Lee, Morrow-Howell, Jonson-Reid, & McCrary, 2010), and that literacy assistance improved general reading achievement (Rebok et al., 2004). Those studies, however, did not evaluate the effectiveness of sustained small group tutoring.

This evaluation, funded by the Corporation for National and Community Service's Social Innovation Fund, is designed to determine whether the ***Experience Corps sustained small group tutoring*** strategy improves children's literacy achievement and social-emotional learning outcomes. Following a two-year implementation evaluation during the 2016-17 and 2017-18 school years, Abt Associates conducted an impact evaluation during the 2018-19 school year. The impact evaluation included three Experience Corps local programs: (1) Generations Inc. (Boston, MA); (2) Read to Succeed Buffalo (Buffalo, NY); and (3) United Way Central Georgia (Bibb County/Macon, GA).

## The Experience Corps Sustained Small Group Tutoring Strategy

The sustained small group tutoring strategy involves tutoring in small groups of up to four students (see Appendix A for a logic model of this strategy). Students meet with a volunteer tutor two to five times per week for approximately 30 minutes, depending on age and need. The group remains together for a sustained period throughout the school year (or the course of the after-school program). AARP recommends that the tutor and small group should meet for at least 35 sessions per school year.

Each local program had a different approach to the session content and structure (see Appendix B for the AARP Foundation's recommended session structure). In Buffalo, skills practice focused on broader reading content, instruction in the five components of reading (phonemic awareness, phonics, fluency,

vocabulary development, and reading comprehension), dialogic reading, guided reading, and writing. In Boston, skills practice focused on phonics, phonemic awareness, phonological awareness, vocabulary, and reading. In Macon, skills practice focused primarily on fluency. Although Macon focused more on reading fluency, there was some fluency instruction in both Boston and Buffalo. Sustained small group tutoring occurred either in a pull-out setting or within the classroom using teacher-provided materials and activities, or materials from an Experience Corps library/toolkit.

## Supporting Literature

Previous research has found significant and positive effects associated with Experience Corps tutoring. The effect of Experience Corps' one-on-one tutoring strategy was examined in a randomized controlled trial of 883 students in grades 1-3 from primarily low-income households in 23 schools across three districts (Lee et al., 2010). That study showed statistically significant increases in passage comprehension and general reading skills of students at risk for reading failure. Another randomized controlled trial of more than 1,000 students in grades K-3 in six urban Baltimore schools, also primarily from low-income households, showed positive and statistically significant classroom effects of a whole-classroom literacy assistance model that incorporated both one-on-one and small group tutoring (Rebok et al., 2004). In 2018, 101 students from three school districts (Minneapolis and St. Paul, MN; and Revere, MA) participated in an evaluation of the impact of Experience Corps tutoring on social-emotional gains. Students showed statistically significant improvements from fall to spring across domains, with the strongest improvements found on the Personal Responsibility, Relationship Skills, and Decision Making scales of the Devereux Student Strengths Assessment (Porowski, de Mars, Kahn-Boesel, & Rodriguez, 2019).

## Impact Evaluation

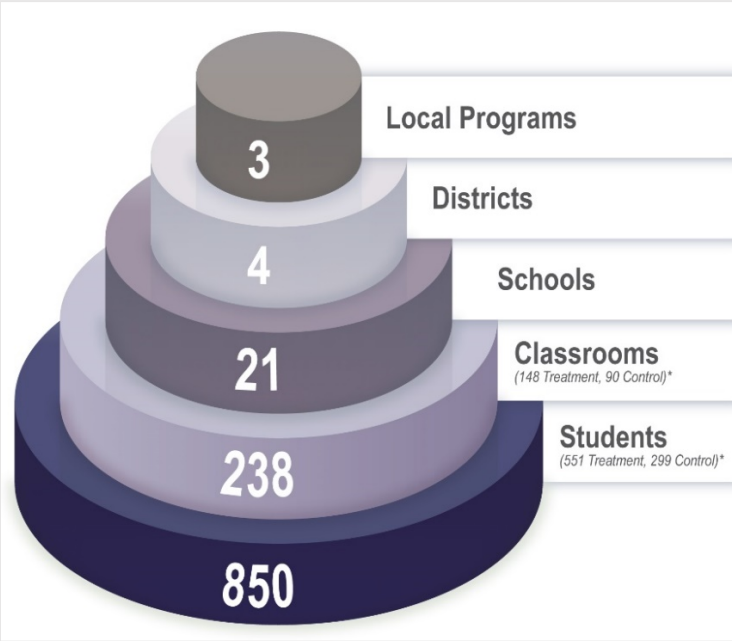
This study examines the impact of Experience Corps' sustained small group tutoring on students' literacy and social-emotional skills, tested with a randomized controlled trial conducted by the Abt evaluation team in the 2018-19 school year. This study builds upon prior experimental research on Experience Corps' one-on-one tutoring by extending the research to the sustained small group tutoring strategy. To build on prior research, this study also explores how any impacts might be contextualized by the nature and quality of program delivery.

Across the three local programs (Boston, Buffalo, and Macon), 244 grade K-3 classrooms in 21 schools were randomly assigned to either receive sustained small group tutoring from Experience Corps tutors (treatment classrooms) or not receive services (control classrooms). One local program (Boston) served schools in two districts: Boston Public Schools and Revere Public Schools. The Buffalo local program served the Buffalo Public Schools, and the Macon local program served Bibb County Schools. Random assignment of classrooms was conducted separately within each school and within each grade level. Because there were often uneven numbers of classrooms at a grade level, the probability of random assignment was adjusted to ensure that at least one classroom was randomly assigned to the control condition in each grade within a given school, and to ensure that Experience Corps could serve as many students as possible during the study.

Within each classroom, all eligible students who consented to be in the study were included in the final analysis. (Six randomized classrooms did not participate in the evaluation, because no students returned consent forms or for other reasons; see Appendix C for details.) The final analytic sample (Exhibit 1)

included 148 treatment classrooms with 551 treatment group students (to receive Experience Corps tutoring) and 90 control classrooms with 299 control group students (to not receive tutoring). At the classroom level, overall attrition was 2.5%, and differential attrition was 0.5%.

**Exhibit 1. Participants in the Final Study Sample**




## Changes to the SIF Evaluation Plan (SEP)

The impact evaluation was originally designed as a three-arm randomized controlled trial comparing the effectiveness of sustained small group tutoring only (Treatment 1) and sustained small group tutoring plus literacy assistance (Treatment 2) to a comparison group that would not receive any Experience Corps services (Control). Based on internal discussions and discussions with grantees, the study team opted to move the study to a two-arm design, focusing on the sustained small group tutoring-only vs. Control comparison as our confirmatory contrast. Literacy assistance was transitioned from a confirmatory to an exploratory contrast, which increased statistical power, simplified the design, eased anticipated recruitment challenges, and more squarely focused the study on the effects of sustained small group tutoring. A copy of the study team's request to change the design is included in Appendix D.

The study team also encouraged local programs to provide literacy assistance to a subset of classrooms that were randomly assigned to deliver sustained small group tutoring. Literacy assistance involves the delivery of supports to meet lesson-related goals at the individual student, small group, or classroom level, so the classroom teacher can focus their attention on the remaining students. Volunteers engaged in literacy assistance provided tutoring and skills practice on an as-needed basis, monitored students when teachers were busy helping other students, and generally provided an additional positive adult presence in the classroom. The inclusion of some literacy assistance supports in some classrooms allowed the study team to measure the value-added of whole-class literacy assistance in addition to the sustained small group tutoring. Because literacy assistance was not randomly assigned to classrooms, this was considered an exploratory analysis. (A detailed description of this impact evaluation's methods and analysis can be found in Appendix C.)

### *Research Questions of Interest*

This report is organized by the following key questions of interest:

- What does the Experience Corps sustained small group tutoring model look like?

- How did students who received Experience Corps sustained small group tutoring compare to students in a control group on literacy achievement and social-emotional learning outcomes?
- Under what conditions is Experience Corps sustained small group tutoring most effective?

### *Impact Evaluation Data Sources*

In fall 2018 and spring 2019, the K-2 battery of the Lexia's **RAPID™ literacy assessment** was administered to both treatment and control group students.[1] The RAPID is a computer-administered adaptive assessment that tests a range of literacy skills.[2] The K-2 battery administered consisted of six subtests: Phonological Awareness, Letter Sounds, Word Reading, Vocabulary Pairs, Following Directions, and Spelling. Of them, students received three or four subtests depending on their grade (Exhibit 2). Because different subtests were administered in different grades, this limited the statistical power of the study.[3]

**Exhibit 2. RAPID Subtest Administration, by Grade and Timing (School Year 2018-19)**

| Outcome Measure | Kindergarten | Grade 1 | Grade 2 | Grade 3 |
|---|---|---|---|---|
| **Phonological awareness** | ✓Fall/Spring | | | |
| **Letter sounds** | ✓Fall | | | |
| **Word reading** | ✓Spring | ✓Fall/Spring | ✓Fall/Spring | ✓Fall/Spring |
| **Vocabulary pairs** | ✓Fall/Spring | ✓Fall/Spring | ✓Fall/Spring | ✓Fall/Spring |
| **Following directions** | | ✓Fall/Spring | ✓Fall/Spring | ✓Fall/Spring |
| **Spelling** | | | ✓Fall/Spring | ✓Fall/Spring |

The RAPID also produces a Reading Success Probability (RSP) score for each student, which refers to the likelihood that a student will achieve grade level success by the end of the year. The RSP is used to identify students who are on track or need additional support.

#### *Literacy Data*

For the impact evaluation, five of the outcome measures in Exhibit 2 were analyzed: Phonological Awareness, Word Reading, Vocabulary Pairs, Following Directions, and Spelling. The letter sounds measure was not included in the analysis because the assessment was only administered at baseline. Sample sizes in these impact analyses differed substantially because not all grades received a given subtest. To explore relationships among findings, implementation variables, and subgroups of interest (e.g., local program), the Abt evaluation team developed a RAPID composite score that represents the average number of standard deviations that each student's RAPID performance diverged from the national norming sample across all available RAPID subtests. Separate composite scores were developed

[1] The subtests of the K-2 battery are different from those in the grade 3 battery (making comparisons across performance scores difficult). In schools where the evaluation team administered the assessment, all selected children regardless of grade received the K-2 battery. In one Boston school, RAPID was administered by school staff; those students received the grade 3 battery.

[2] The RAPID's reliability coefficients for the five assessments used in the impact evaluation range from .85 to .94.

[3] Assuming an intraclass correlation of .10, alpha level of .05, power of .80, and 49% of variance explained by covariates, the minimum detectable effect sizes for each outcome were .43 for Phonological Awareness, .26 for Word Reading, .21 for Vocabulary Pairs, .26 for Following Directions, and .32 for Spelling. These differences underscore the challenges introduced in detecting effects as the RAPID subtests do not cover the full sample.

for each student at pretest and posttest, and these scores ensured that the most data possible could be used in the exploratory analyses.[4]

#### *Social-Emotional Learning Data*

In November 2018 and again in May 2019, teachers in the treatment and control classrooms were invited to complete the eight-item Devereux Student Strengths Assessment (**DESSA-mini**) to assess the social-emotional behaviors of their students participating in the evaluation. The DESSA-mini is a subset of the larger 72-item DESSA assessment and includes eight SEL competencies: self-awareness, self-management, social awareness, relationship skills, personal responsibility, decision making, goal-directed behavior, and optimistic thinking. The DESSA-mini yields a T-score for each student. The T-score is a standard score, scaled to have a mean of 50 and a standard deviation of 10.[5]

### *Implementation Data Sources*

In spring 2019, teachers, principals, tutors, and program directors were invited to complete **surveys** to share their perspectives and experiences on the implementation of the Experience Corps program and its hypothesized effects. Additionally, the evaluation team conducted 67 Experience Corps **tutoring session observations** during spring 2019. Altogether, the evaluation team conducted 18 observations in Boston, 18 observations in Buffalo, and 31 observations in Macon. Local programs were asked to provide access to a more experienced tutor and less experienced tutor in each school where observations took place. Each tutoring session observation was guided by an observation tool that guided observers in tracking each element of the Experience Corps session, including session characteristics; communication between the volunteer tutor and student; skill building (with a focus on tracking fluency practice); review, reflection, and quality talk; and tutor read-aloud. See Appendix E for a copy of the observation tool.

The team was also provided with **session log data** recorded by tutors after each tutoring session, which provided valuable information about the dosage of Experience Corps tutoring that students received.

### *Analytic Approach*

Student outcomes were analyzed in two-level linear models that tested the effect of Experience Corps on the five RAPID subtests as well as the DESSA-mini, accounting for the nesting of students within classrooms. Covariates in each model included a student's pretest measure for a given assessment, baseline RSP score, race/ethnicity, gender, grade level, special education status, and English learner status. To account for the probability of randomization, each randomization block (i.e., grade level within a school) was included as a dummy variable in the analysis.

[4] The RAPID composite is an exploratory metric that represents the average z-score of all RAPID subtests. Z-scores were derived from mean performance scores and standard error of measurement reported in Appendix B of the RAPID Technical Manual (Foorman, Petscher, & Schatschneider, 2019).

[5] The median internal consistency of the DESSA-mini is reported to be .92 (RAND Education Assessment Finder, 2018)

# What does the Experience Corps sustained small group tutoring model look like?

To provide context for the impact of Experience Corps sustained small group tutoring, the evaluation team collected information about the characteristics of the tutoring: the nature of the intervention, including dosage; adherence to the recommended session structure; tutoring quality; tutor characteristics; and additional classroom supports for students struggling to read. This section of the report provides descriptive information on those characteristics and findings related to how those characteristics differed across local programs.

### Key Findings

- On average, students received 34.2 sessions and 15.8 hours of tutoring in 2018-19.
- Classroom observations and tutor surveys revealed that on average across local programs, instruction, quality talk, and read-aloud portions of the session were delivered approximately in the same proportions as the Experience Corps recommended session structure; however, there was substantial variation between local programs in small group tutoring delivery and the content delivered.
- Macon tutors were observed to adhere most closely to the recommended session structure, and Macon received the highest quality ratings for its tutoring.
- Macon had the smallest average group size (2.7 students), followed by Buffalo (3.2 students) and Boston (3.5 students).

## Descriptive Information

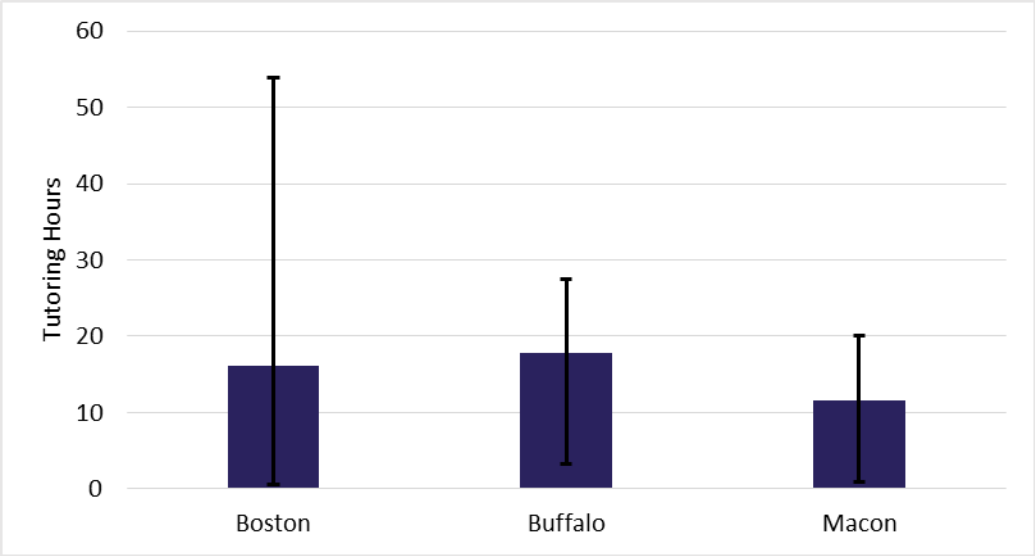
### *Tutoring Dosage*

Session log data collected by the AARP Foundation and reported by Experience Corps tutors capture tutoring dosage. Among the students in the treatment group who received at least one Experience Corps small group session (*n*=527), the average number of sessions received was 34.2 over the school year. The average number of sessions per week was 1.7. Moreover, the average number of hours of Experience Corps sustained small group tutoring received by students was 15.8.[6]

The number of sessions an average student in the treatment group received differed substantially by local program. Exhibit 3 below shows the average number of Experience Corps sustained small group tutoring hours received by students in each local program, as well as the range (shown with vertical bars). The overall average number of hours was lower in Macon than in Boston and Buffalo (11 vs. 16 and 18 hours, respectively), and the range of hours across students was also smaller in Macon. Among students in the treatment group who received at least one tutoring session during the year, 58% of these students received at least 35 sessions (45% in Boston, 84% in Buffalo, and 36% in Macon).

[6] Twelve (12) students in the control group received one or more Experience Corps tutoring sessions, and 24 students in the treatment group did not receive any Experience Corps tutoring sessions. These 36 students are not included in the dosage numbers presented here. These students' outcomes were analyzed based on the group to which they were originally randomly assigned.

**Exhibit 3. Number of Experience Corps Sustained Small Group Tutoring Hours Received by Students**




## *Tutoring Session Structure*

### *Location*

Most (75%) of the observed tutoring sessions took place outside of the classroom (30% in resource rooms, 27% in libraries, 12% in hallways, 6% in computer labs/cafeterias/other places). The remainder (25%) occurred within the students' classroom. The AARP Foundation does not expect tutoring sessions to occur in the same location across programs or even across schools, as location is often predicated on the available physical space within school buildings and instructional scheduling within grades. The AARP Foundation, however, acknowledges that location can make a difference for students, and each setting has both benefits and disadvantages to learning. Volunteers who deliver tutoring during instructional times within the classroom may battle additional distractions but benefit from easier access to the classroom teacher and instructional materials. In pull-out settings such as the library or a resource room, tutors may deliver support in quieter, more focused environments. Such settings, however, may create challenges to teacher-tutor coordination.

At the local program level, there were differences in where Experience Corps sustained small group tutoring occurred. Of the 18 observed sessions in Boston, 78% took place within the classroom, compared to 6% of the 18 observations in Buffalo and 7% of the 31 observations in Macon.

### *Length and Focus*

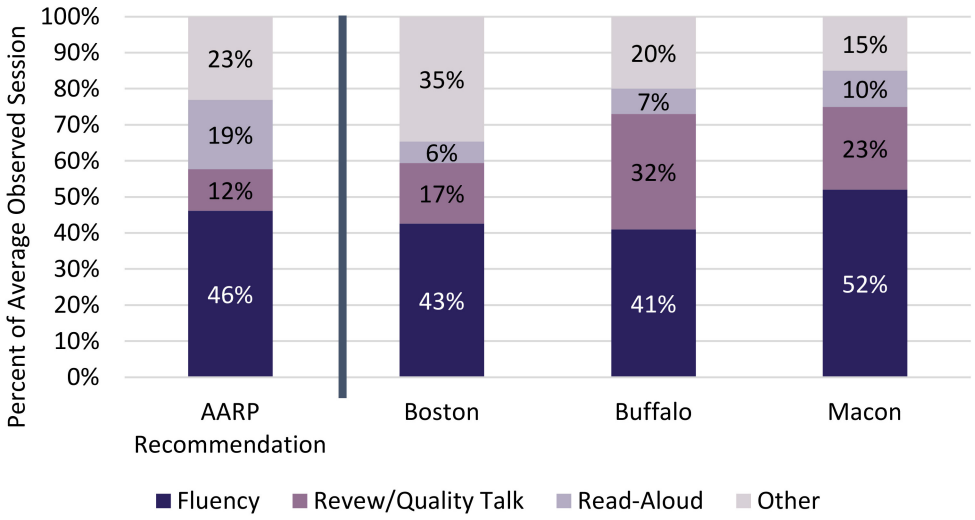
The AARP Foundation recommends that Experience Corps sustained small group tutoring sessions be approximately 30 minutes long.[7] Results from the tutor survey revealed that most (63%) of the observed tutoring sessions were between 15 and 29 minutes in length; 31% were between 30 and 45 minutes, and the remainder were very short (3% were less than 15 minutes) or very long (3% were longer than 45 minutes). The average tutoring session was 27 minutes long, and it focused on skill building for 46% of the session, on quality talk for 24% of the session, on tutor read-aloud for 8% of the session, and on

[7] The recommended session structure calls for a minimum of 26 minutes in each session. The evaluation team found that 32 of the 67 observed sessions (48%) lasted less than 26 minutes (50% in Boston, 39% in Buffalo, and 52% in Macon).

other content for 22% of the session. Activities categorized as "other" included social talk, transition time, setting up/cleaning up materials, behavior management, and gift distribution (e.g., stickers, candy).

In its *Structured Session Guide* (Appendix B), the AARP Foundation recommends that 46% of the session be dedicated to skill building (fluency for those following the Experience Corps model), 12% to review/quality talk, and 19% to tutor read-aloud; the remaining 23% of the session might focus on opening and closing components. Exhibit 4 shows how the average observed percentage of time dedicated to these components differed by local program. All three local programs, on average, came close to the recommended time dedicated to core instruction, but they diverged in time dedicated to quality talk and tutor read-aloud.

**Exhibit 4. Observed Average Distribution of Experience Corps Sustained Small Group Tutoring Time, by Component and Local Program**




The tutor survey also asked tutors about how they used their small group time (Exhibit 5, on the following page). The overall distributions were similar to those observed. Tutors reported that they spent most of their time (52%) on skill building activities. They spent an average of 21% of their time on quality talk and 15% on tutor read-aloud.

**Exhibit 5. Tutor Reported Average Distribution of Experience Corps Sustained Small Group Tutoring Time, by Component and Local Program**

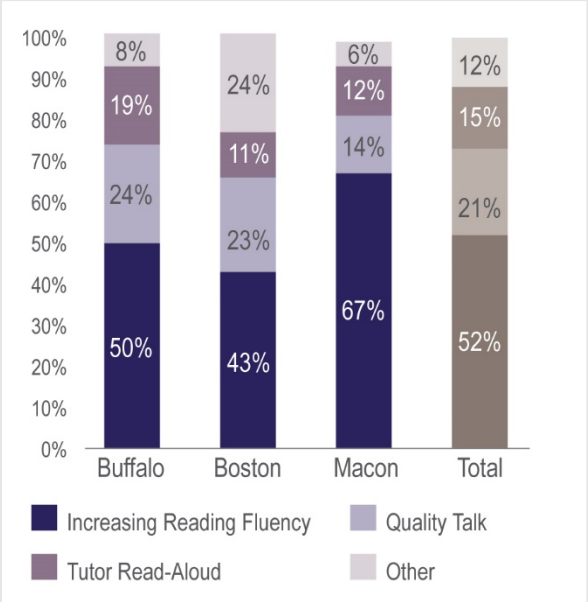


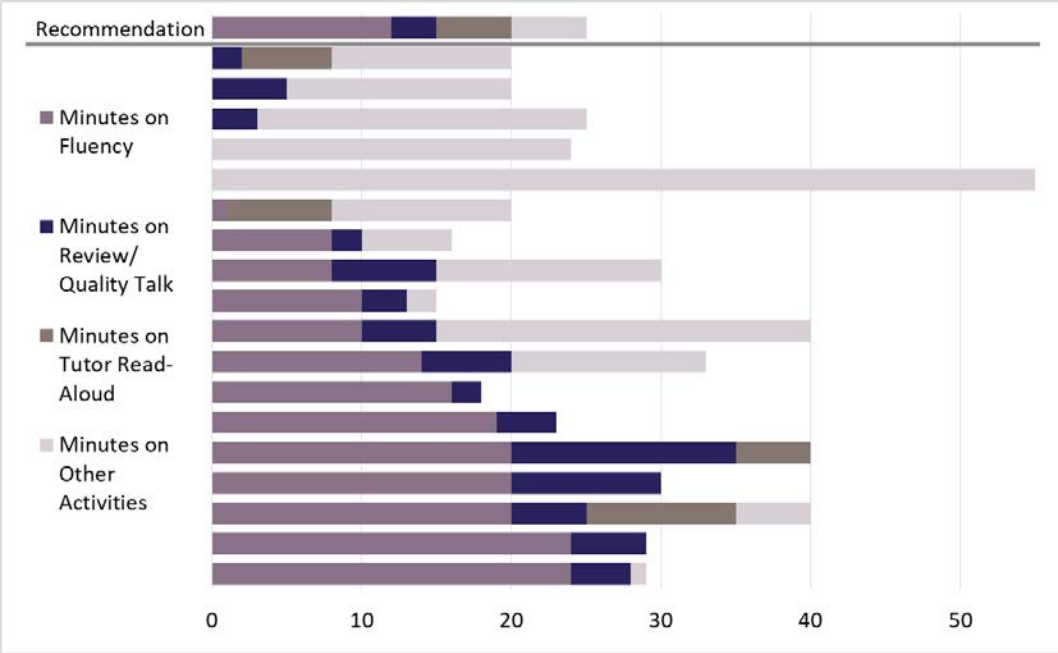

Exhibit 6 (beginning on the following page) shows the number of minutes in each observed tutoring session that was focused on the four key session components: skill building (fluency or other core instruction), review/quality talk, tutor read-aloud, and other activities. The top bar in each graph highlights the AARP Foundation Experience Corps recommendations.

The number of minutes dedicated to each key component varied substantially across sessions as well as varied in fidelity to the AARP Foundation recommendations. Boston generally varied from the recommendations the most, and tutor read-aloud was absent from most observed sessions there. Macon most closely aligned to the recommendations. Buffalo fell between the other two. As noted previously, the content of the skill building component differed by local program.

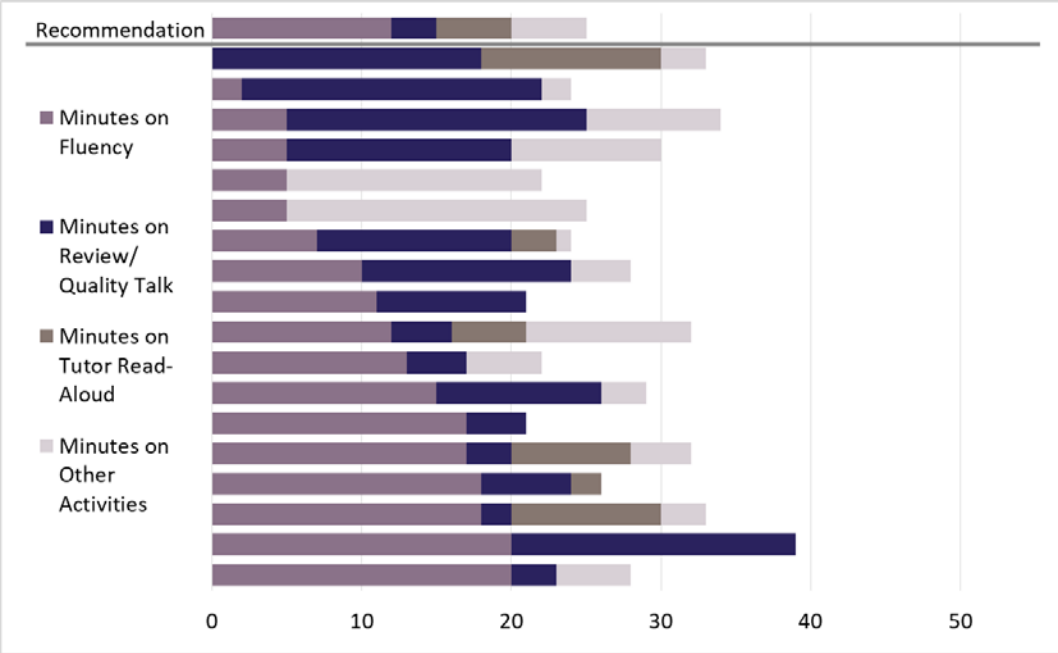
As noted in the introduction, each of the three local programs had a different skill focus. Macon focused primarily on fluency. Buffalo focused on the five components of reading along with dialogic reading, guided reading, and writing. Boston focused on the five components of reading (phonics, phonemic awareness, phonological awareness, vocabulary, and reading). The observation tool used by the study team focused its observations of skill building portions of the tutoring sessions on fluency practice, as Experience Corps emphasized fluency in their structured session guidelines.

**Exhibit 6. Distribution of Observed Time per Session of Experience Corps Sustained Small Group Tutoring, by Local Program**

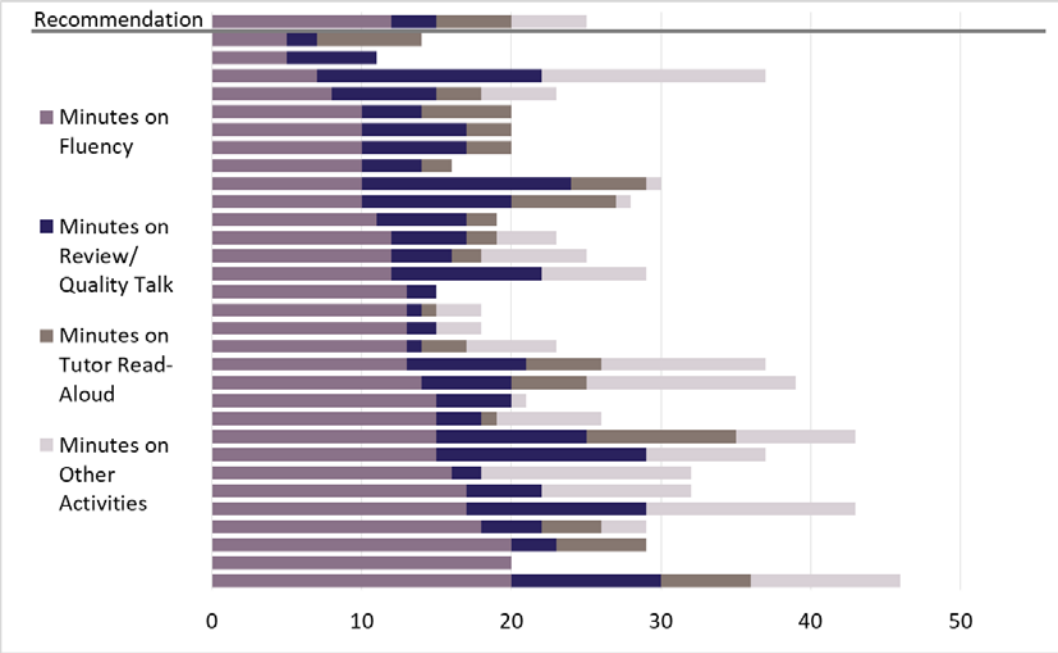
***Boston***




***Buffalo***



***Macon***

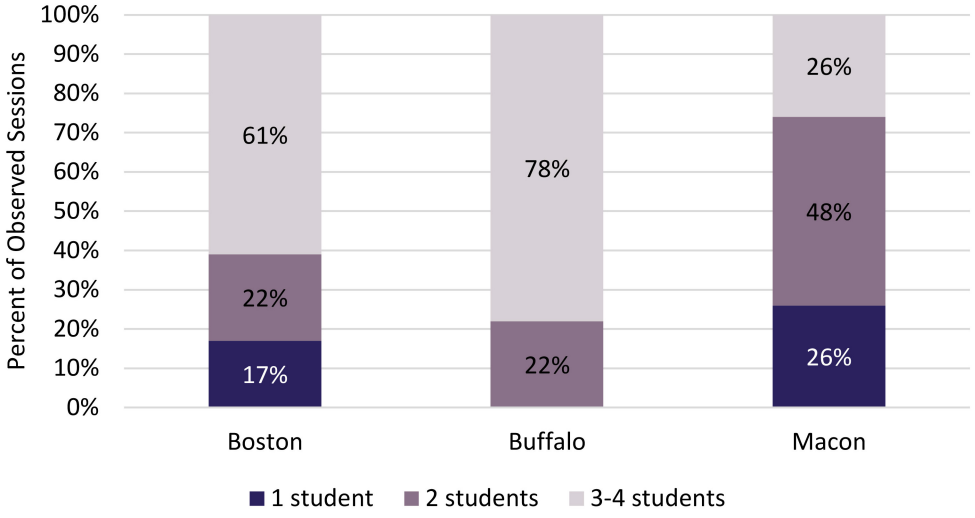



### *Size*

As defined by the AARP Foundation's Experience Corps standards, a small group should include no more than four students. Among the sessions observed, 34% involved a group of two students and 49% involved a group of three or four students. However, 16% of sessions involved a single student. Because these observations were conducted at a single point in time, it is possible that the 11 observations of one-on-one tutoring (3 in Boston, 8 in Macon) were simply the result of other students being absent on the day of the observation. Group size differed substantially by local program, as seen in Exhibit 7. Nearly three-quarters of observed sessions in Macon included only one or two students, whereas observed sessions in Boston and Buffalo largely comprised three or four students.[8]

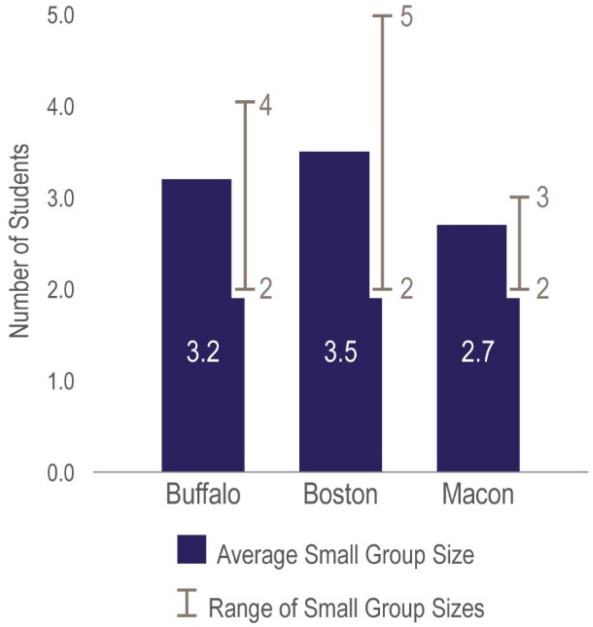
[8] Some observed sessions involved only one student (3 in Boston and 8 in Macon). Though observations only captured one session at one point in time, the sampling plan for the observations should not have resulted in a biased sample, or a sample of sessions that was not necessarily representative of typical day-to-day tutoring activities. In order to select the sample for observations, local programs provided a less-experienced and a more-experienced tutor from each school.

**Exhibit 7. Observed Small Group Size, by Local Program**




The observations provide a comprehensive and dynamic view of a small subset of sessions; surveys were used to capture information from a larger number of tutors about their experiences. Tutors were surveyed about the size of their small groups during a typical session, and their responses differed somewhat from the observed sessions. In general, tutors reported working with between two and five students at a time (Exhibit 8). Whereas all of the tutors in Macon and Buffalo reported working with between two and four students, 9% of Boston tutors reported working with more than four students in their group during a regular week. On average, group sizes were the largest in Boston and the smallest in Macon, according to the tutor survey.

**Exhibit 8. Small Group Size Reported by Tutor, by Local Program**

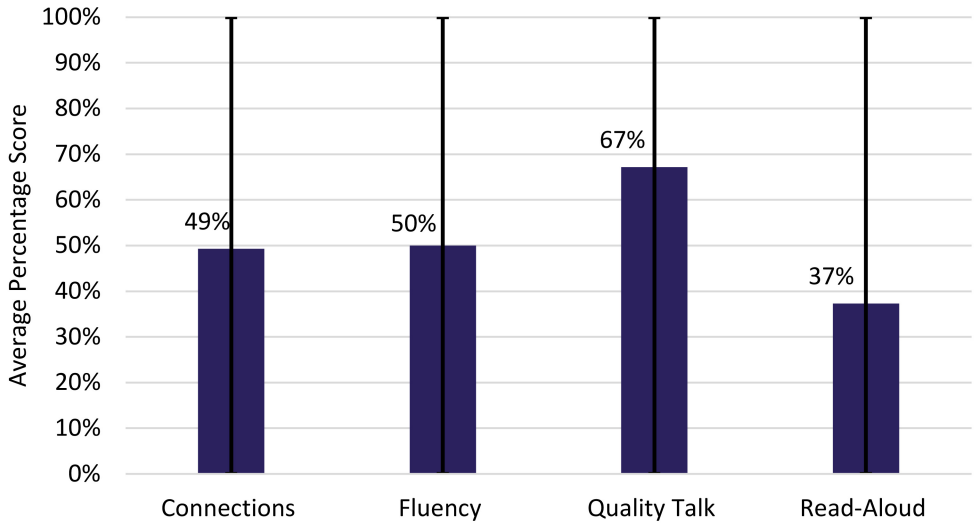


### *Tutoring Session Quality*

Tutoring session quality was scored by the evaluation team based on elements in the observation protocol in the following categories: connections, fluency, quality talk, and tutor read-aloud. A score was assigned in each of the four categories based on the percentage of elements within each category that were observed. Exhibit 9 below shows that, for example, in the skill building category, the average tutoring session dedicated time to half (50%) of the observation protocol's fluency-related elements. The vertical bars show the observed range across sessions. The range for every category was 0%-100%, meaning at least one observed session never included an element from that category, and at least one observed session included all elements from that category.

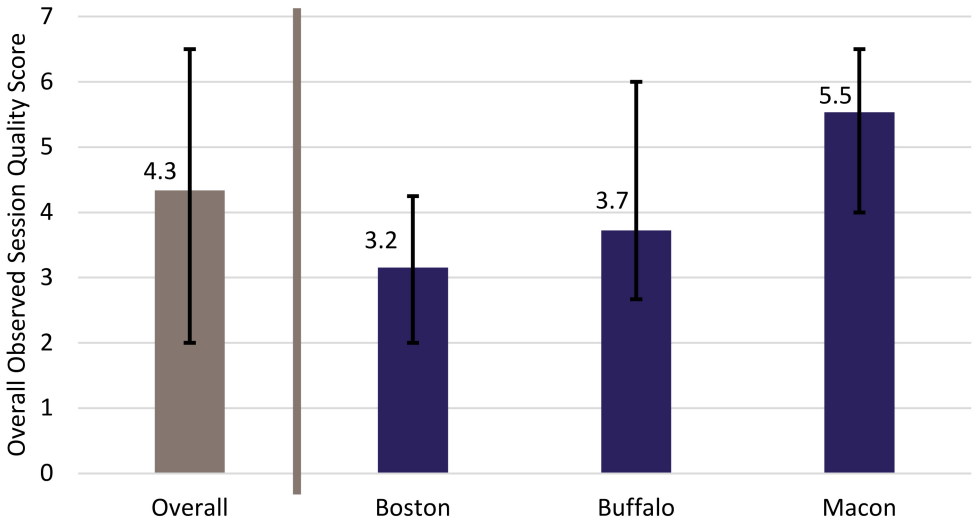
The evaluation team also determined an overall quality score based on these categories, with more weight given to fluency skill building than to the other three categories. Overall session quality was calculated by assigning 1 point within each of the three other categories (connections, quality talk, and tutor read-aloud) if at least one element in the category was observed, plus 1 point for each fluency element observed.

**Exhibit 9. Observed Tutoring Session Quality, by Category (All Local Programs)**




The overall average quality of tutoring sessions, as seen in Exhibit 10, was 4.3 on a scale of 0-7 (standard deviation = 1.4). The exhibit shows the average overall observed quality by local program, with the range of observed quality represented with vertical bars. Using this scale, the highest observed session quality was in Macon. Boston and Buffalo had lower quality ratings, as the session quality rating was weighted more heavily for sessions that focused on fluency. Macon focused the skill building portion of their structured session more heavily on fluency, while there was somewhat less of a focus on fluency in Boston and Buffalo, as noted above.

**Exhibit 10. Average Observed Tutoring Session Quality, by Local Program**

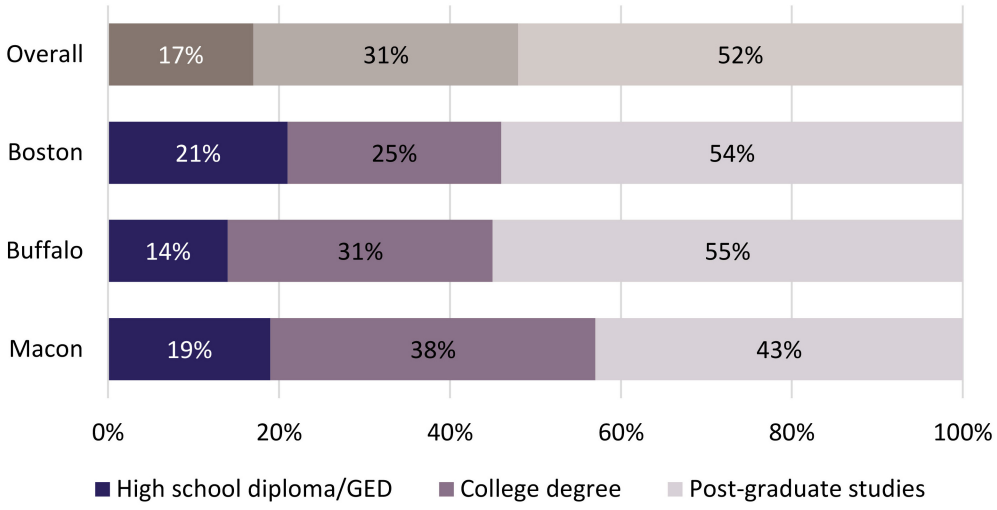



## Tutor Characteristics

Across the 87 tutor surveys[9], 74% were returning tutors from the previous year (83% in Boston, 69% in Buffalo, and 71% in Macon). The average number of years of experience with Experience Corps was 2.2 across the sample of responding tutors (range: 1 year to 5 years). In Boston the average was 2.6, in Buffalo the average was 2.4, and in Macon the average was 1.9. The majority of tutors had a college degree or higher (Exhibit 11), and there was little variation by local program.

Tutor characteristics were not analyzed in relationship to student gains due to difficulties linking tutors to classrooms consistently in the datasets.

**Exhibit 11. Tutor Education, by Local Program**

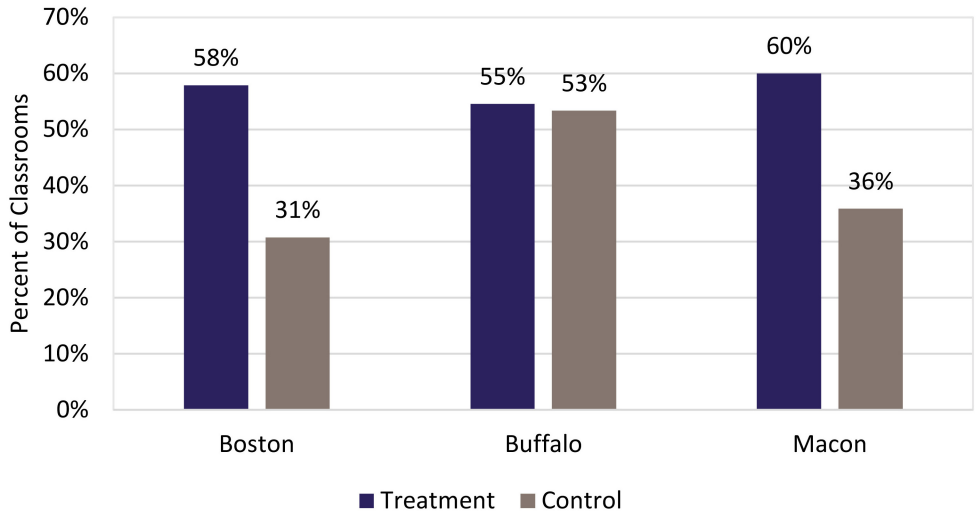



[9] This represents 70% of all tutors in study schools.

## Additional Literacy Supports

In the teacher survey, teachers were asked to identify additional literacy and social-emotional supports present in their classrooms, above and beyond typical classroom instruction—and above and beyond Experience Corps, for teachers in the treatment classrooms. Half of the responding[10] teachers in the study reported having at least one additional formal literacy support program associated with a Response to Intervention (RTI) framework in their classroom[11] serving students who participated in the study (58% of teachers in treatment classrooms and 39% of teachers in control classrooms). Some 30% of responding teachers in the study reported having at least one formal social-emotional program in their classroom[12] serving students who participated in the study (37% of teachers in treatment classrooms and 21% of teachers in control classrooms). Exhibit 12 shows the percentages of classrooms using these formal programmatic literacy supports by condition and local program.

**Exhibit 12. Teachers Reporting Using Additional Literacy Supports in Their Classroom, by Condition and Local Program**




[10] These numbers represent only 68% of the classrooms in this study.

[11] For teachers in treatment classrooms, in addition to Experience Corps sustained small group tutoring. Examples of reported formal literacy support programs are Response to Intervention, Lexia Core, and TutorMate.

[12] For teachers in treatment classrooms, in addition to Experience Corps sustained small group tutoring. Examples of reported formal social-emotional programs are Restorative Circles, Second Step, and Positive Behavior Interventions and Supports (PBIS).

# How did students who received Experience Corps sustained small group tutoring compare to the control group on literacy achievement and social-emotional learning outcomes?

## ✓ Experience Corps students made gains at the same rate as the control group, despite starting the year lower, on average, in literacy achievement.

At the end of the school year, no statistically significant differences were observed between students who received Experience Corps sustained small group tutoring and students who did not. Effect sizes for each contrast are included in Exhibit 13 below, which represent the number of standard deviation units that separate the treatment and control groups. For reference, the U.S. Department of Education's What Works Clearinghouse (WWC) defines a "substantively important" effect as at least .25.[13]

Despite random assignment of treatment and control classrooms, there were some significant group differences in student skills at the beginning of the school year. Although effect size differences at baseline were within What Works Clearinghouse standards of .25 standard deviation units, baseline differences were larger than expected, ranging from −.18 (RAPID Spelling) to +.18 (DESSA-mini). Baseline differences were statistically significant on the RSP, suggesting that students in Experience Corps had systematically lower literacy skills at baseline and were predicted to have a lower chance of achieving end-of-year proficiency.

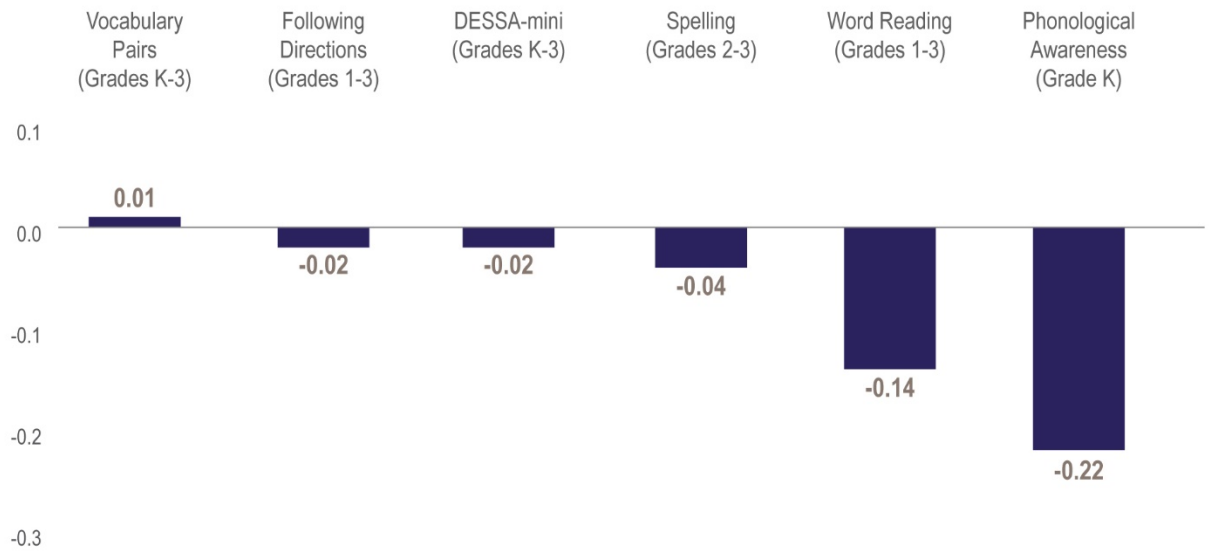
### Key Findings

- Experience Corps sustained small group tutoring did not have a statistically significant impact on student literacy or social-emotional skills, when compared to the control group.
- Both Experience Corps students and control group students had a statistically significant gain in social-emotional skills, as measured by the DESSA-mini.
- Students in treatment classrooms began the year with somewhat lower literacy skills but higher social-emotional skills than students in the control classrooms.
- Experience Corps students had greater baseline-posttest gains on RAPID scale scores than the control group, but these findings did not control for demographics or the probability of random assignment.
- Exploratory analyses reveal that tutoring dosage was significantly related to gains in literacy and social-emotional skills.
- Students in classrooms that received Experience Corps literacy assistance in addition to Experience Corps sustained small group tutoring had significantly lower gains on RAPID Vocabulary Pairs than students who received Experience Corps sustained small group tutoring alone.

[13] The What Works Clearinghouse has since dropped this designation, but the WWC still uses .25 standard deviations as the maximum allowable baseline difference for quasi-experimental designs.

Even though impact estimates for all measures except vocabulary pairs were negative, they were not significantly different from zero. Although the effect sizes for phonological awareness may appear to be substantial, readers should not over-interpret this result, as only kindergarteners were included in that analysis and the unadjusted outcomes indicate an equally strong effect size *in favor of* the Experience Corps group (Exhibit E.2). There are a number of possible reasons why impacts were not observed, several of which are explored further in this report.[14]

**Exhibit 13. Treatment-Control Group Differences at the End of the School Year**




Note: Findings reported as effect sizes (Hedges' *g*).

## ✓Experience Corps students improved their literacy and social-emotional skills over the course of the year.

Exhibit 14 below displays the average covariate-adjusted means at each time point by condition for the RAPID Vocabulary Pairs, RAPID Word Reading, and DESSA-mini student assessments. Although these charts clearly show an upward trajectory in literacy and social-emotional skills among the treatment group, they also show similar upward trajectories for the control group.[15]

[14] The RAPID RSP score was not included as an outcome variable in the analysis because it is designed as a diagnostic metric to predict the probability of reading success by the end of the school year. The RSP was used as a baseline covariate, however.

[15] In order to achieve an effect size of .25 (and assuming the same standard deviation in both the treatment and control groups), the average RAPID Word Reading scale score for the treatment group would have needed to be 593.99; the treatment group's Vocabulary Pairs scale score would have needed to be 518.56; and the DESSA-mini average t-score would have needed to be 56.93.

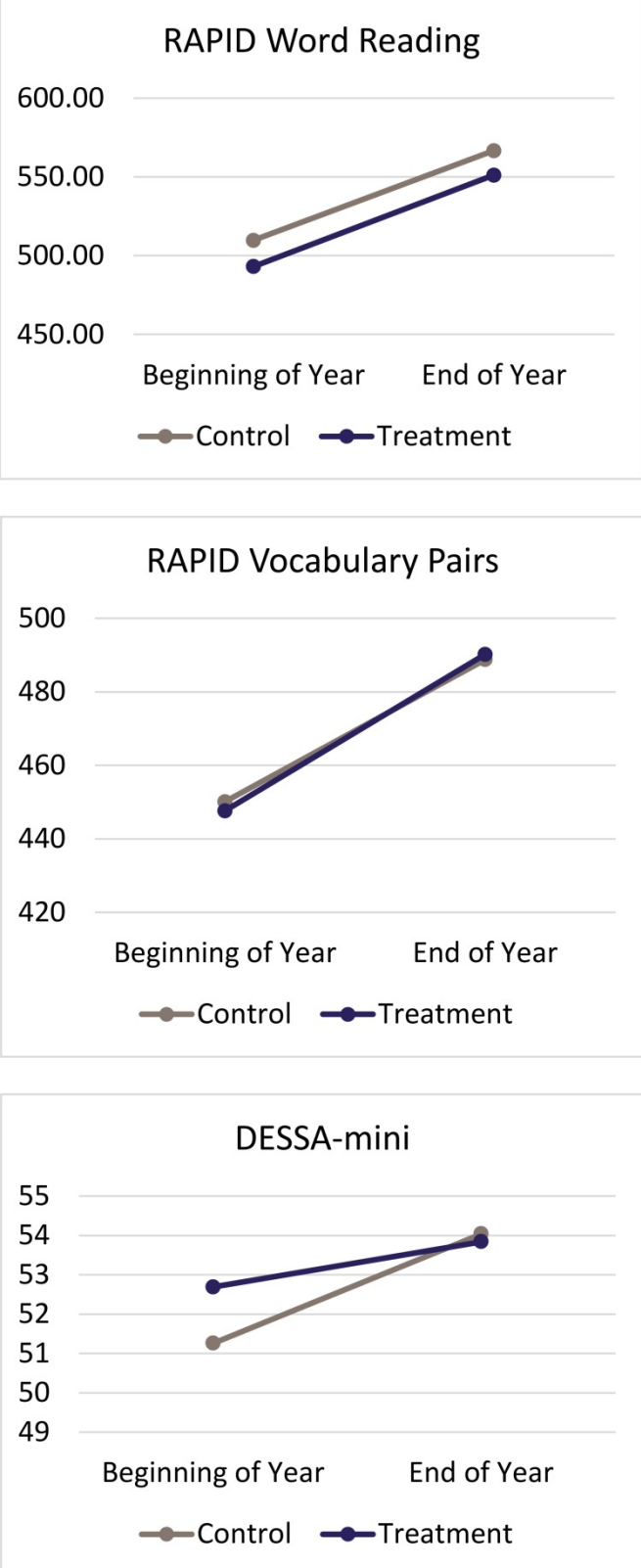
Teacher survey data revealed that teachers in both treatment and control classrooms reported observing improvements in student skills over the course of the year. Approximately 85% of the teachers reported observing improvements in literacy skills, and around 50% of teachers reported observing improvements in students' social-emotional skills. More teachers in classrooms reported seeing improvements in treatment student behavior than teachers in control classrooms (55% vs. 46%).

An exploratory analysis of gain scores (i.e., baseline-posttest change in RAPID assessment scores) reveals that Experience Corps students on average had larger pre-post gains than students in the control group (Exhibit 15 on the following page). Moreover, a larger proportion of Experience Corps students improved on the RAPID by more than 100 points from baseline to posttest than the control group in word reading, phonological awareness, and spelling. In analyses controlling for demographics and the probability of assignment, students in both the treatment group and the control group significantly improved their literacy and social-emotional skills over the course of the school year, making similar gains. The lack of statistically significant gains for the Experience Corps group relative to the control group contradicts previous positive findings reported for Experience Corps' one-on-one tutoring (Lee et al., 2010).

✓ No significant differences between Experience Corps and control classrooms emerged from benchmark data.

Each district provided the evaluation team with benchmark data that was used both to validate RAPID findings and to explore measures of fluency and other literacy skills not covered in the RAPID.

**Exhibit 14. Average Adjusted Literacy Growth over Time**



Because each local program's benchmark assessments were different and scored on different scales, these benchmark assessments were analyzed separately (Star 360® in Macon, Dynamic Indicators of Basic Early Literacy Skills [DIBELS®] in Buffalo, DIBELS Oral Reading Fluency [DORF]-Accuracy in Boston). As such, findings from the benchmark assessments could not be pooled together and did not provide sufficient statistical power to detect statistically significant differences between the treatment and control groups, even when controlling for baseline measures and grade level.

## ✓ Experience Corps sustained small group tutoring dosage was significantly related to gains in literacy and social-emotional skills.

In exploratory analyses,[16] the number of hours of Experience Corps sustained small group tutoring that a student received emerged as a statistically significant predictor of overall RAPID performance ($p$=.040). Dosage was also significantly related to gains in social-emotional skills; students who received more hours of Experience Corps sustained small group tutoring also had larger gains on the DESSA-mini ($p$=.017).

**Exhibit 15: Profile of Baseline-Posttest Growth on RAPID Scale Scores, by Condition**

| | Condition | n | Average Growth | Percent of Students with Pre-Post RAPID Growth (scale score points) | | | |
|---|---|---|---|---|---|---|---|
| | | | | Negative or No Change | +1 to +50 | +51 to +100 | + More than 100 |
| **Word Reading** | EC | 344 | 60.7 | 25% | 40% | 12% | 23% |
| | Control | 169 | 57.7 | 22% | 42% | 16% | 20% |
| **Phonological Awareness** | EC | 127 | 232.0 | 4% | 13% | 5% | 79% |
| | Control | 73 | 218.9 | 10% | 8% | 10% | 73% |
| **Vocabulary Pairs** | EC | 469 | 41.6 | 33% | 27% | 19% | 22% |
| | Control | 240 | 39.1 | 33% | 29% | 15% | 23% |
| **Following Directions** | EC | 342 | 20.3 | 44% | 25% | 13% | 18% |
| | Control | 168 | 18.8 | 43% | 23% | 15% | 18% |
| **Spelling** | EC | 218 | 72.5 | 15% | 39% | 16% | 30% |
| | Control | 113 | 61.5 | 15% | 50% | 15% | 20% |

[16] Dosage was not included in the primary (confirmatory) impact analyses because it is an endogenous covariate. Endogenous covariates are covariates in statistical models that are obtained after baseline and are influenced by group status. These covariates can lead to biased results, but may be used for exploratory purposes.

## Relationship between Outcomes and Implementation

Within the treatment group, the evaluation team examined the relationship between student gains in literacy and social-emotional skills and characteristics of the intervention.

### ✓ Adherence to the Experience Corps sustained small group tutoring session structure was not related to treatment gains.

The length of observed Experience Corps sustained small group tutoring sessions; the number of students in a tutoring session; and the average number of session minutes spent on fluency, quality talk, or tutor read-aloud were not significantly related to student gains on any measure.

### ✓ Sustained small group tutoring session quality was not significantly related to treatment gains.

The overall Experience Corps sustained small group tutoring session quality measure, which was based on observations of a subsample of tutors and aggregated to the school level, was not significantly related to a student's gains on any measure; however, this relationship was positive and approached significance on the RAPID Vocabulary Pairs subtest ($p$=.093).

### ✓ Additional literacy supports (including literacy assistance) were not significantly related to treatment gains.

The presence of classroom-level literacy supports in addition to Experience Corps sustained small group tutoring provided to students in the treatment classrooms was not significantly related to student gains on any measure.

In addition to sustained small group tutoring, some treatment classrooms opted to provide Experience Corps whole-class literacy assistance. In the literacy assistance strategy, tutors provide support to teachers to reinforce literacy skills or concepts introduced in daily lessons. The teacher assigns the tutor to work with a small group, an individual student, or an entire class while the teacher works with other students in the classroom. Literacy assistance is employed to meet a specific lesson-related goal, rather than individual student goals. Across treatment classrooms, students in classrooms that received Experience Corps literacy assistance in addition to Experience Corps sustained small group tutoring had significantly lower gains on RAPID Vocabulary Pairs ($B$=−33.23, SE=15.64, $p$=.036) than did students in Experience Corps classrooms that did not also provide literacy assistance. There was no significant relationship between the presence of literacy assistance and scores on the RAPID composite[17] or the DESSA-mini.

[17] The RAPID composite is an exploratory metric that represents the average z-score of all RAPID subtests. Z-scores were derived from mean performance scores and standard error of measurement reported in Appendix B of the RAPID Technical Manual (Foorman, Petscher, & Schatschneider, 2019).

# Under what conditions is Experience Corps sustained small group tutoring most effective?

## Exploratory Analyses

Experience Corps' sustained small group tutoring did not have a statistically significant effect on students' literacy or social-emotional skill outcomes as measured by the RAPID and DESSA-mini, respectively. However, the primary impact analyses and additional exploratory analyses can provide insights on the situations where the Experience Corps model may work best.

In the following section, exploratory findings are shared about the differential effects of Experience Corps by local program, student subgroups, RAPID baseline data, and the presence of other literacy supports in the classroom.

### ✓ Macon had stronger average outcomes than Buffalo and Boston.

The exploratory analysis—which examined outcomes of students' average RAPID subtest scores transformed to z-scores—revealed that Macon had the strongest findings, with the difference in outcomes between Macon and Boston being statistically significant (see Appendix F). There may be a connection between Macon having the strongest overall findings and the highest observed session quality scores. Another explanation may be that Macon had the best balance between treatment and control group RSP scores at baseline.

Macon's average categorical RSP score[18] in the fall of 2018 was 1.22 for both its treatment group and its control group. In contrast, Buffalo's RSP scores at baseline were 1.27 for the treatment group and 1.39

### Key Findings

- Experience Corps sustained small group tutoring was most effective in Macon.
- Macon also had the best treatment/control balance in RAPID baseline RSP scores.
- There was no systematic pattern of gender, English learner, grade level, race/ethnicity, or special education status being significantly related to RAPID performance.
- The presence of additional formal literacy support programs in classrooms did not change the overall impact results.
- The amount of time a child spent in Experience Corps tutoring over the year was a significant predictor of positive results on the RAPID.
- Although statistically significant findings were not observed on the RAPID or the DESSA-mini, teachers in all three local programs reported improvements in students' literacy skills.
- The positive relationship between dosage and outcomes, coupled with the positive findings from Macon, suggests a systematic relationship between the Experience Corps model delivery and outcomes.
- More research is needed to unpack these findings.

[18] The RSP's categories are coded as: 1=Low probability of reading success (RSP<=30), 2=Moderate probability of reading success (RSP between 31 and 69), and 3=High probability of reading success (RSP>=70).

for the control group, and Boston's RSP scores at baseline were 1.40 for the treatment group and 1.62 for the control group. Although RSP baseline scores were used as a covariate in the impact models, further work is needed to disentangle the implications of these imbalances. After all, RSP scores are a validated predictor of reading at grade level by the end of the school year. Boston's lower average baseline RSP scores in the treatment group (relative to the control group) may have largely explained its students' lower performance in the exploratory analyses.

## ✓ No significant patterns emerged between student outcomes and student characteristics.

In the primary multi-level model findings, there was no pattern of statistically significant relationships between student outcomes and student characteristics such as gender, English learner status, grade level, or baseline performance. However, some results were significant:

- Boys had stronger performance on the RAPID Word Reading subtest than did girls.
- Black, Hispanic, and Asian students all had significantly lower performance on the RAPID Phonological Awareness subtest in kindergarten than did White kindergartners.
- Students in lower grades (especially kindergarten) had lower performance on the RAPID Vocabulary Pairs subtest than did third-graders.

Moreover, English learners had weaker performance on the exploratory analyses that combined and averaged RAPID subtest scores, yet this pattern did not emerge in the primary HLM findings.

## ✓ No significant patterns of findings emerged by RAPID subtest.

There was no pattern of significant program impact on any of the subtests of the RAPID assessment or on any individual item from the DESSA-mini teacher ratings.

## ✓ Additional literacy supports did not affect results.

Using teacher survey responses about literacy supports provided to students in the evaluation, the evaluation team created a four-group condition variable, which is shown in Exhibit 16 along with associated classroom numbers.

**Exhibit 16. Presence of Additional Literacy Supports, by Experimental Condition**

| Condition | Control Classrooms | Treatment Classrooms | Total |
|---|---|---|---|
| **No formal literacy support program** | 41 | 0 | 41 |
| **Non-Experience Corps formal literacy support program only** | 26 | 0 | 26 |
| **Experience Corps sustained small group tutoring only** | 0 | 39 | 39 |
| **Experience Corps sustained small group tutoring + additional formal literacy support program** | 0 | 53 | 53 |
| **Missing information** | 23 | 56 | 79 |

This new four-group condition was not significantly related to student gains on any measure; students in classrooms who received Experience Corps sustained small group tutoring in combination with a formal literacy support program did not make stronger gains than students in other classrooms, including classrooms with no formal literacy support program at all.

Most survey respondents not only reported observing strong improvements in student progress, but also attributed them to the presence of Experience Corps.

For example, Exhibit 17 below shows that at least 75% of teachers in each local program reported improvements in literacy skills among Experience Corps students. Fewer than 10% of teachers in each local program reported that Experience Corps students did not improve literacy skills. These data were drawn from a teacher survey that asked teachers about general gains in literacy, so it is possible that teachers are noticing gains in student performance that may not be measured by the RAPID.

Exhibit 18 below shows that teachers overwhelmingly reported that students received a sufficient number of Experience Corps tutoring sessions. It is possible that as teachers reported strong student improvement in literacy, those changes were also taking place in non-Experience Corps classrooms.

**Exhibit 17. Teacher Report of Student Improvement in Literacy Skills (Experience Corps Students Only)**

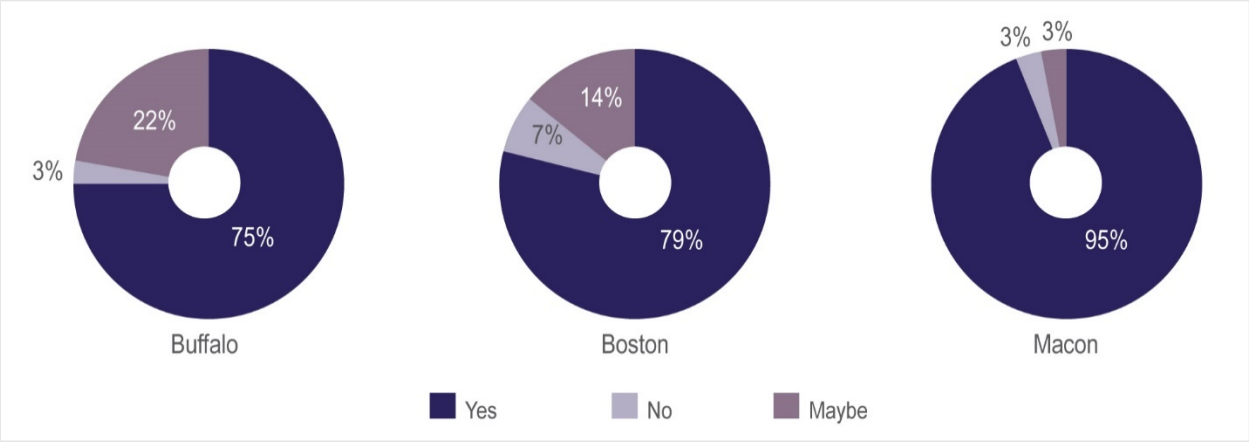
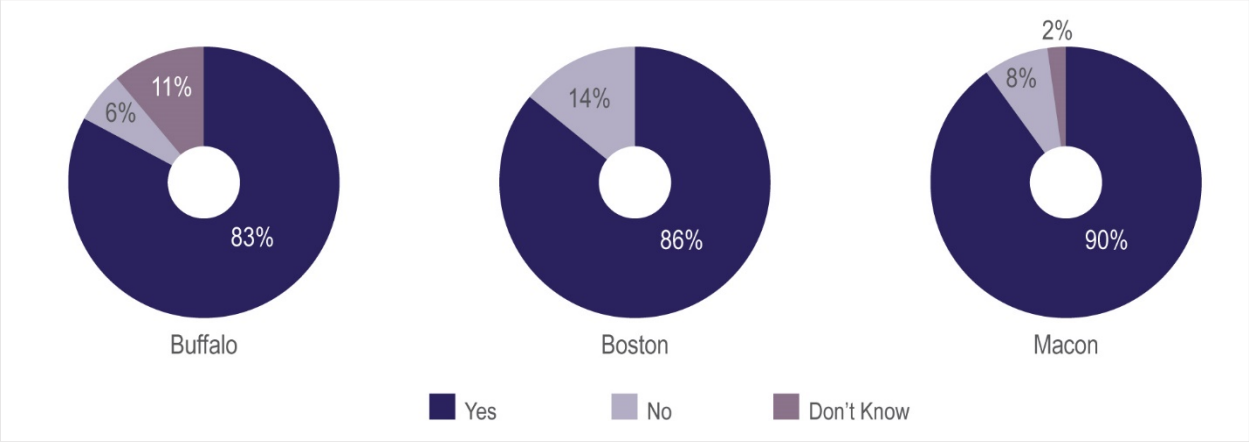



**Exhibit 18. Teacher Report of Whether Students Received a Sufficient Number of Tutoring Sessions (Experience Corps Students Only)**



# Conclusion

Some contributing factors may at least partially explain why Experience Corps sustained small group tutoring did not have statistically significant impacts on literacy and social emotional skills:

- There is a possibility that the evaluation team experienced "unhappy" randomization. That the RSP scores were substantially lower for the treatment group than the control group in Buffalo and Boston means there may have been a systematic source of bias introduced into the study in the student selection process. For example, it is possible that students in the treatment group received more follow-up encouragement to return consent forms to participate in the study than the control group students. If this was the case, it may indicate that the control group had more intrinsic motivation than the treatment group (i.e., control students didn't need multiple follow-ups). This hypothesis is purely conjecture by the study team, as the consent process was administered locally. Still, it is evident from these figures that the treatment and control groups had some unexpectedly large baseline differences for a randomized controlled trial. Baseline differences were still within the .25 standard deviation standard set by the What Works Clearinghouse for quasi-experimental designs.
- Because there is no posttest composite score from the RAPID, all impact analyses were conducted separately by individual RAPID subtest, which limited the statistical power of the study.
- The positive findings in Macon may be explained by Macon having much smaller group sizes than Boston and Buffalo. These findings from Macon may help explain the positive findings from one-on-one tutoring found by Lee et al. (2010).
- The positive findings in Macon also may have been the result of Macon having adhered most closely to Experience Corps' recommended session structure, and it having the highest session quality scores.
- It is possible that the RAPID did not precisely measure the skills being taught. The Experience Corps model is focused on fluency, but the RAPID does not have a formal fluency measure. Thus, it is possible that the precision of measurement was limited by the lack of a fluency measure. However, exploratory analyses of fluency benchmark assessment data (e.g., DIBELS, Star 360) did not yield positive findings in favor of the treatment group.

The positive relationship uncovered in the exploratory analyses between dosage and outcomes, coupled with the positive findings from Macon, indicates that there may be some systematic relationship between the Experience Corps model delivery and outcomes. More research is needed to unpack these findings.

In response to these findings, the AARP Foundation has taken a number of affirmative steps:

- **In fall 2020, Experience Corps will be delivered using a one-on-one tutoring strategy only**. This move away from sustained small group tutoring began after Abt presented these evaluation findings to the AARP Foundation, and accelerated in response to the COVID-19 pandemic.

Experience Corps tutoring will move to a virtual format for the fall 2020 semester, which would have made sustained small group tutoring much more difficult to implement.

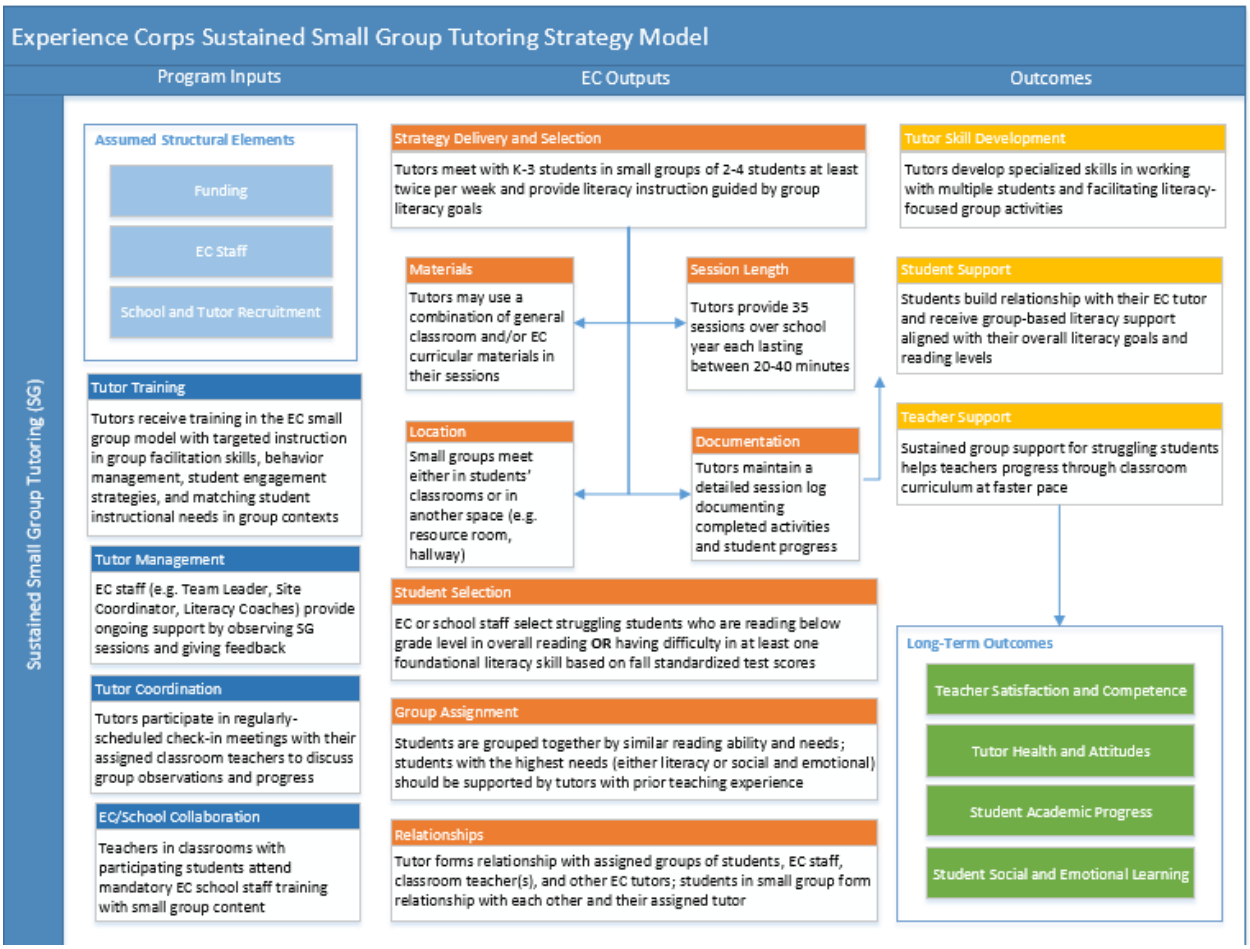
- **The AARP Foundation no longer supports literacy assistance**. The exploratory findings from this study did not establish a positive relationship between literacy assistance and literacy outcomes, nor social-emotional learning outcomes. Since the AARP Foundation is moving to a virtual tutoring model in fall 2020, there is no longer a need to support literacy assistance. In response to these findings, the AARP Foundation has made this a permanent change.
- **The AARP Foundation has redoubled efforts to encourage adoption of its new structured session guidelines**. Both Buffalo and Boston have committed to adopt the new structured session guidelines starting in the fall of 2020.

In addition to applying lessons learned from this study, the AARP Foundation is supporting efforts for continued learning about how its structured session can influence fluency outcomes. Starting in the fall of 2020, AARP Foundation Experience Corps local programs will be asked to collect and report student-level fluency data from standardized assessments instead of relying on teacher progress reports. This will provide for more rigorous measurement of student progress and set the stage for continued learning for the foreseeable future.

# Appendix A: The Experience Corps Sustained Small Group Tutoring Logic Model

As a part of the **program inputs**, tutors receive pre-service and in-service trainings that guide their activities and interactions with students and school staff. Tutors also learn how to interpret assessment data and communicate with teachers about student progress; and receive oversight from Experience Corps staff, who provide feedback, coaching, and support. Teachers may attend Experience Corps trainings, meet with their classroom's Experience Corps tutor, and provide materials/curriculum for tutors.

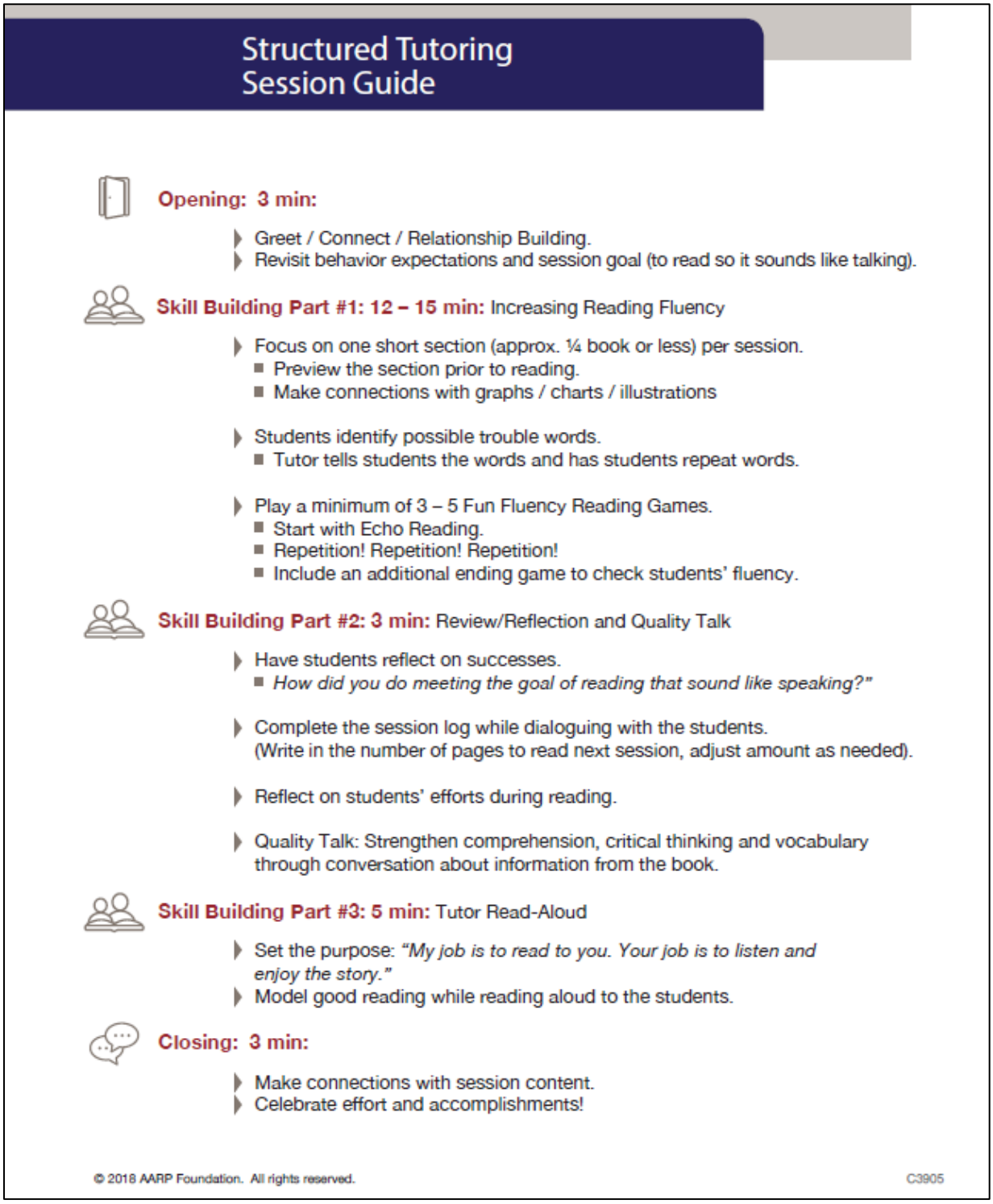
The logic model also includes **Experience Corps outputs**. School and/or Experience Corps staff identify classrooms to receive program services, and teachers in those classrooms select students to receive tutoring. The strategy delivery includes tutors providing academic and social-emotional support to a group of students over the course of the school year. Student starting points in leveled readers is determined based on student assessment outcomes, and dosage is tracked through session logs. The Experience Corps outputs also include building relationships among teachers, students, Experience Corps staff, school staff, and tutors.



Four mediating **outcomes** are described in the Experience Corps program: tutor skill development, student support, positive classroom/school climate, and teacher support. Tutors develop skills related to their tutoring, including literacy activities and behavior management. Students may build a relationship with the Experience Corps tutor, who provides them consistent academic and social-emotional support. The additional adult presence in the classroom may help improve classroom/school climate, and teachers can receive direct assistance as a part of the Experience Corps program model.

The program tracks six **long-term outcomes** for students, teachers, and tutors. Students gain positive social-emotional skills and improve their reading through their participation in the Experience Corps program. Teachers may receive extra support in the classroom, contributing to their overall job satisfaction. Teachers may also gain capacity to differentiate instruction to more students when their students participate in the program. Tutors may show improved physical, cognitive, and social health, in addition to improved attitudes regarding their self-efficacy and connection to the community.

# Appendix B: The Experience Corps Structured Session



## Structured Tutoring Session Guide

**Opening: 3 min:**

- Greet / Connect / Relationship Building.
- Revisit behavior expectations and session goal (to read so it sounds like talking).

**Skill Building Part #1: 12 – 15 min:** Increasing Reading Fluency

- Focus on one short section (approx. ¼ book or less) per session.
  - Preview the section prior to reading.
  - Make connections with graphs / charts / illustrations
- Students identify possible trouble words.
  - Tutor tells students the words and has students repeat words.
- Play a minimum of 3 – 5 Fun Fluency Reading Games.
  - Start with Echo Reading.
  - Repetition! Repetition! Repetition!
  - Include an additional ending game to check students' fluency.

**Skill Building Part #2: 3 min:** Review/Reflection and Quality Talk

- Have students reflect on successes.
  - *How did you do meeting the goal of reading that sound like speaking?"*
- Complete the session log while dialoguing with the students. (Write in the number of pages to read next session, adjust amount as needed).
- Reflect on students' efforts during reading.
- Quality Talk: Strengthen comprehension, critical thinking and vocabulary through conversation about information from the book.

**Skill Building Part #3: 5 min:** Tutor Read-Aloud

- Set the purpose: *"My job is to read to you. Your job is to listen and enjoy the story."*
- Model good reading while reading aloud to the students.

**Closing: 3 min:**

- Make connections with session content.
- Celebrate effort and accomplishments!



C3905

# Appendix C: Methods and Analysis

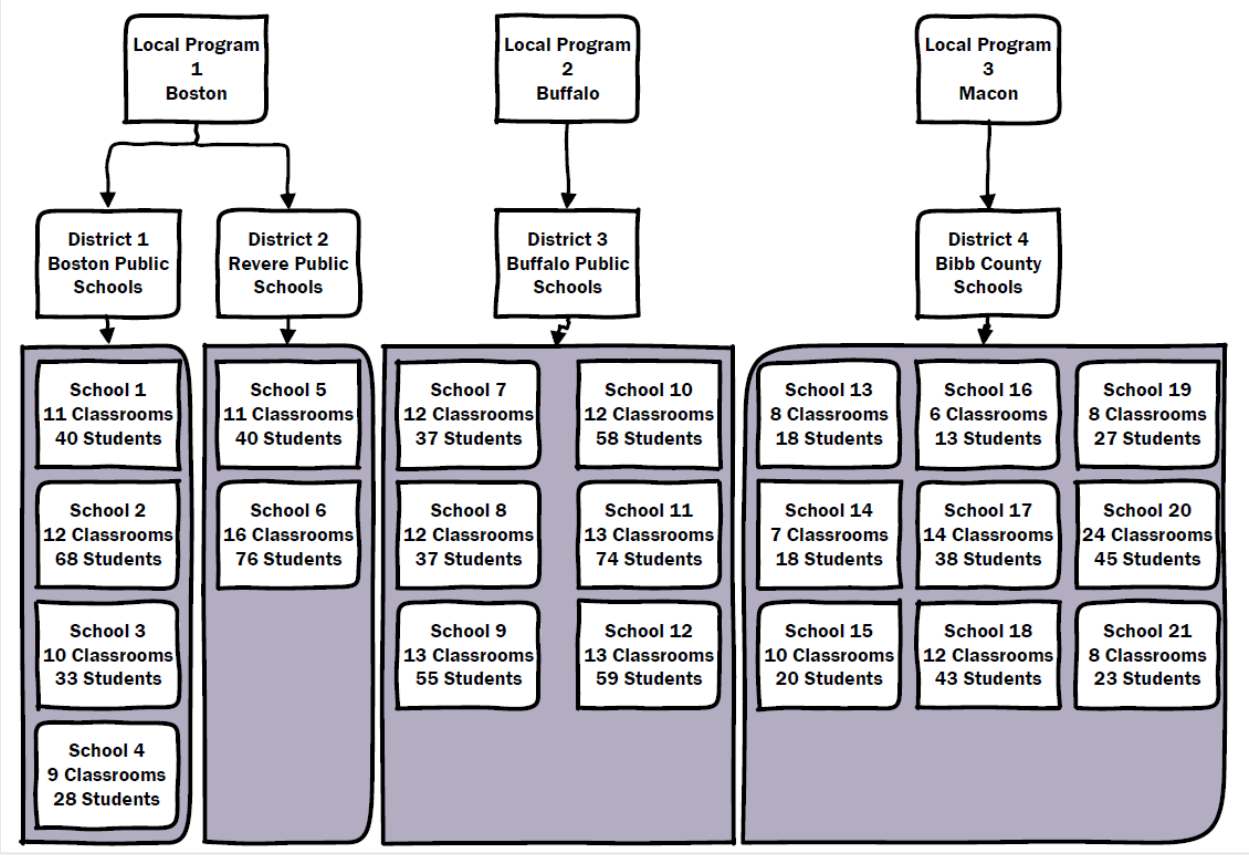
## Methods

### *Study Design*

#### *Sample Recruitment and Randomization*

The sample for the evaluation was a five-level nested sample with 850 students nested within 238 classrooms, classrooms nested within 21 schools, schools nested within four school districts, and districts nested within three Experience Corps local programs (SIF sub-grantees). Exhibit C-1 shows the general sample organization. Participating *classrooms* across grades K-3 were randomly assigned to either the treatment condition (receiving Experience Corps tutoring) or the control condition. Not all four grade levels were represented in each participating school.

**Exhibit C-1. Sample Organization**




#### *Local Programs*

At the end of the 2017-18 school year, the evaluation team conducted an evaluability assessment to inform the selection of sub-grantees to participate in the 2018-19 impact and implementation evaluations. This evaluability assessment was grounded in the implementation data collected from each sub-grantee local program.

In May 2018, the evaluation team applied the evaluability criteria and assigned points to each local program for each of the criteria. These criteria were designed to capture the optimal conditions in place to support a randomized controlled trial. After calculating the total scores across the criteria for each local program, the evaluation team selected the top three highest-scoring local programs to participate in the impact evaluation. The three local programs selected to participate in the evaluation were Generations, Inc. (Boston, MA), Read to Succeed Buffalo (Buffalo, NY), and United Way of Central Georgia (Macon, GA). Buffalo and Macon each served one school district, Buffalo Public Schools and Bibb County Schools, respectively. Boston served two districts, Boston Public Schools and Revere Public Schools.

### *Schools/Principals*

Schools were eligible to be in the evaluation if the school district granted permission for the research and principals consented for their schools to be in the evaluation. To participate, classrooms in a school could not be completely saturated with Experience Corps tutors, meaning at least one classroom per grade needed to be designated as a business-as-usual control classroom with no access to Experience Corps services. If classrooms in a school were saturated with Experience Corps tutors, the tutors could be reallocated to other schools in the district during the study year. Exhibit C-2 shows the number of schools served by each Experience Corps local program that were recruited to participate in the impact evaluation.

**Exhibit C-2. Participating School Numbers, by Local Program**

| Local Program | Number of Participating Schools |
|---|---|
| **Boston** | 6 |
| **Buffalo** | 6 |
| **Macon** | 9 |
| **Total** | 21 |

### *Classrooms/Teachers*

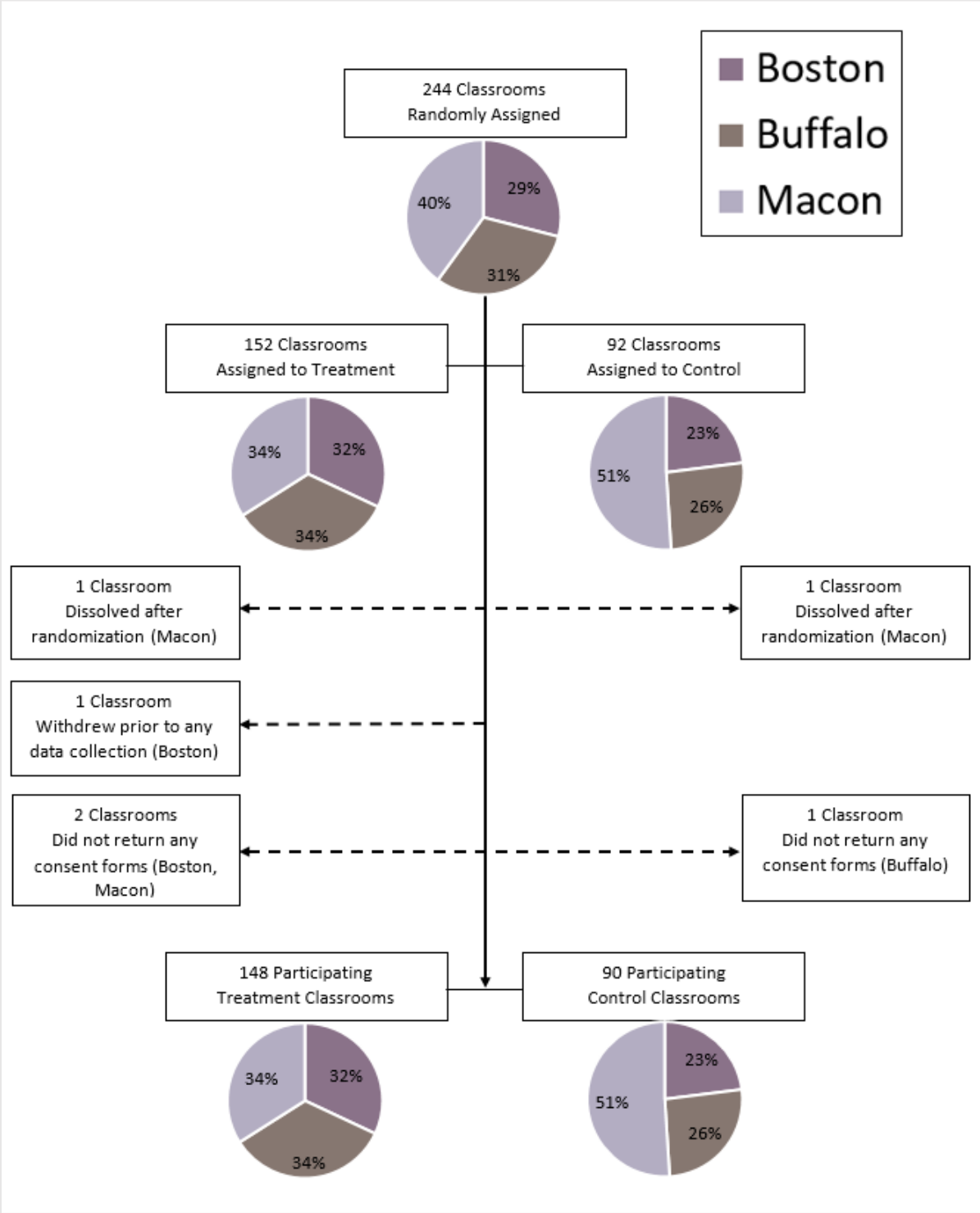
A participating school's classrooms were eligible to participate in the study if they did not exclusively serve students in special education, though blended classrooms were eligible. Local programs were encouraged to ensure that each participating grade within a school had at least two classrooms in the study (in which case, one of the two would be randomly assigned to the treatment condition and the other to the control condition).[19]

In the 2018-19 school year, after students had been assigned to teachers using standard school practices, the evaluation team followed randomization procedures to assign classrooms within a school and a grade either to receive Experience Corps sustained small group tutoring (treatment condition) or to not receive any Experience Corps services (control condition). The probability of random assignment was adjusted based on the number of classrooms that needed to be randomized within each school and grade. The decision for how many classrooms within a given school and grade to randomly assign to the treatment condition was made in consultation with local EC staff.

[19] Three schools had some grades that did not have classrooms in both experimental conditions. One additional school had classrooms in both experimental conditions, but the only control classroom in one grade level did not have any consented, eligible children.

Exhibit C-3 below shows the number of classrooms that were randomly assigned to each condition (and the distribution by local program). As depicted in the exhibit, two control and four treatment classrooms were randomized but did not participate in the evaluation, because they either dissolved, withdrew, or had no students return consent forms. That left 238 study classrooms in the analysis sample.

**Exhibit C-3. Classroom-Level CONSORT Chart**



*Students*

Students were eligible to participate in the study if they were eligible to receive Experience Corps tutoring and if their parent(s) consented to their being studied. Eligibility of students was determined by the teacher and/or any other procedure that schools typically used to allocate students to tutoring, such as scores on reading assessments. For instance, if a school typically assigned students to receive an Experience Corps tutor when the child's DIBELS® phonemic awareness scores were between 10 and 20, then teachers in control classrooms and treatment classrooms in that same school were able to use those same DIBELS criteria to readily identify students to be randomly assigned to the control group or treatment group.

Within treatment classrooms, teachers were asked to ensure that eligible students randomly selected to receive Experience Corps sustained small group tutoring did receive it. Teachers in treatment classrooms could also include other students in the tutoring as they wished. And teachers in control classrooms were asked to ensure that their students did not receive Experience Corps services, but they were not prohibited from providing other types of tutoring to their students.

The evaluation team attempted to obtain parental consent for all students in eligible classrooms. First program and/or school staff determined which students were eligible to receive Experience Corps services. Then the team randomly selected up to eight students per classroom to participate in the study from among students who both had parental consent and were Experience Corps eligible.[20] The number of students included in the evaluation is shown in Exhibit C-4 and student demographics in Exhibit C-5.

**Exhibit C-4. Student Numbers, by Condition and Local Program**

| Local Program | Control Group | Treatment Group | Total |
|---|---|---|---|
| **Boston** | 91 | 194 | 285 |
| **Buffalo** | 98 | 222 | 320 |
| **Macon** | 110 | 135 | 245 |
| **Total** | 299 | 551 | 850 |

**Exhibit C-5. Student Demographics, by Condition**

| Demographic | Control Group | Treatment Group | Overall |
|---|---|---|---|
| Race/Ethnicity | | | |
| Black/African-American | 53% | 42% | 46% |
| Hispanic/Latinx | 15% | 26% | 22% |
| White/Caucasian | 17% | 19% | 18% |
| Female | 49% | 56% | 53% |
| Special education status | 5% | 9% | 8% |
| English learner status | 23% | 19% | 20% |

Note: Special education status meant the student had an Individualized Education Plan and/or 504 plan. English learner status was specific to the year the data were collected. Both designations were determined by the district.

[20] Rarely did a classroom have more than eight consented, eligible children; thus, random selection of children was not often needed. Occasionally, a local program or district requested that the evaluation involve only a limited number of children (sometimes only two children) from a given classroom; in those cases, random selection of consented, eligible children was applied.

# Measurement

## *Impact Data Sources*

### *RAPID*

In fall 2018 and spring 2019, the K-2 battery of Lexia's RAPID™ literacy assessment was administered to treatment and control students in participating classrooms in each of the study districts. The evaluation team administered the RAPID to students in all nine Macon schools, all six Buffalo schools, and both of the Revere schools. Of the four schools in the Boston Public School District, the team administered the RAPID in two schools in the fall and one in the spring. Otherwise, students were administered the RAPID by their teachers as a part of their regular benchmark assessments.

The subtests administered as part of the K-2 battery are different from those administered as part of the grade 3 battery (making comparisons across performance scores difficult). For that reason, in schools where the evaluation team administered the RAPID, all selected children regardless of grade received the K-2 battery. The grade 3 students In the Boston school where RAPID was administered by school staff, grade K-2 students received the K-2 battery, but grade 3 students received the grade 3 battery.

Students typically complete the RAPID in one class period (15-25 minutes for the K-2 battery) using a computer or tablet connected to the internet. The RAPID consists of six subtests, of which five were administered for the impact evaluation. In each subtest, its first five items (two easy, one medium, and two hard items) are designed to quickly measure student ability, and how students respond to them determines the subsequent test items presented to the student.

Depending on their grade, students received three or four subtests among those focused on phonological awareness, letter sounds, word reading, vocabulary pairs, following directions, and spelling (Exhibit C-6). All batteries at all grades consisted of a mix of teacher-led tasks (i.e., requires one-on-one administration) and student-led tasks (i.e., the student works independently). Because some tasks were student-led, test administrators were able to use a cascade-style administration to assess students (i.e., once the first student has completed the teacher-led tasks, the administrator can begin working with a second student as the first student completes the student-led tasks).

**Exhibit C-6. RAPID Subtest Administration by Grade and Timing**

| Subtest | Kindergarten | Grade 1 | Grade 2 | Grade 3 |
|---|---|---|---|---|
| **Phonological Awareness** | ✓Fall/Spring | | | |
| **Letter Sounds** | ✓Fall | | | |
| **Word Reading** | ✓Spring | ✓Fall/Spring | ✓Fall/Spring | ✓Fall/Spring |
| **Vocabulary Pairs** | ✓Fall/Spring | ✓Fall/Spring | ✓Fall/Spring | ✓Fall/Spring |
| **Following Directions** | | ✓Fall/Spring | ✓Fall/Spring | ✓Fall/Spring |
| **Spelling** | | | ✓Fall/Spring | ✓Fall/Spring |

Students in all grades completed at least one word recognition subtest (Phonological Awareness, Letter Sounds, Word Reading, or Spelling) and at least one academic language subtest (Vocabulary Pairs or Following Directions).

1. In the **Phonological Awareness** subtest, students hear words that have been broken into parts and must blend the parts into one word and say it aloud. This subtest is teacher-led (assessment administrators score in real time).

2. In the **Letter Sounds** subtest, students view letters and consonant diagraphs and say the corresponding sounds. This subtest is teacher-led (assessment administrators score in real time).
3. In the **Word Reading** subtest, students read words aloud that they see on the screen. This subtest is teacher-led (assessment administrators score in real time).
4. In the **Vocabulary Pairs** subtest, students are presented with three words and/or pictures and must select the two that go together. This subtest is student-led (scored by the computer).
5. In the **Following Directions** subtest, students follow audio directions to move and/or select pictures on a screen. This subtest is student-led (scored by the computer).
6. In the **Spelling** subtest, students hear words and spell them using the computer keyboard. This subtest is student-led (scored by the computer).

Scoring of the test is automatic and yields norm-referenced performance scores ranging from 0 to 1000 for individual RAPID tasks; percentile ranks can be calculated from these performance scores. The scoring also produces a Reading Success Probability (RSP) score, which predicts the likelihood that a student will achieve grade level success by the end of the year, that is used to identify students who are on track or need additional support. According to Lexia, an RSP of 70 or higher means that the student has high likelihood of reaching end-of-year grade level success. An RSP between 31 and 69 means moderate likelihood; and an RSP of 30 or lower means low likelihood.

### *DESSA-mini*

In November 2018 and again in May 2019, teachers in the sample were invited to complete surveys about the social-emotional behaviors of their students participating in the study (in both the treatment and control groups).

Survey questions were taken from the Devereux Student Strengths Assessment (DESSA). The full DESSA consists of 72 items, eight of which can be used to rate student social-emotional competence. Added together, these eight items produce a DESSA-mini score. The majority of teachers in this evaluation rated student skills across only those eight items. However, there were a subset of teachers in two Revere schools participating in a concurrent study in which they were asked to complete the full DESSA for each of their participating students. For those students, only their responses to the eight DESSA-mini items were pulled out for analysis in this evaluation.

The DESSA-mini yields a T-score for each student. The T-score is a standard score, scaled to have a mean of 50 and a standard deviation of 10. These T-scores can range from 28 to 72, and differences in T-scores from pretest to posttest have the same interpretation throughout the range. For example, a five-point change in T-score from 35 to 40 represents the same magnitude difference as a five-point change in T-score from 60 to 65.

### *Data Provided by School Districts*

Each participating school district provided the evaluation team with a range of student data, including attendance, behavior, demographic data, and a variety of benchmark assessments used. Exhibit C-7 provides an overview of the most popular benchmark assessments for each district.

**Exhibit C-7. District-Collected Benchmark Data Measures**

| Measure | Boston | Buffalo | Macon |
|---|---|---|---|
| **RAPID™** | ✓ (3 schools) | | |
| **MAP Reading Fluency™** | ✓ (1 school) | | |
| **Fountas & Pinnell™** | ✓ (1 school) | | |
| **DIBELS®** | ✓ (2 schools) | ✓ (all schools) | |
| **Renaissance Star 360®** | | | ✓ (all schools) |

Notes on benchmark assessments:

- School districts included Beginning of the Year (BOY), Middle of the Year (MOY), and End of the Year (EOY) data for the various **reading benchmarks** administered within each school district throughout the year:
    - All Buffalo schools and two Boston schools reported scores for the DIBELS, but some of the standard subtests varied within each program. Buffalo schools reported Phoneme Segmentation Fluency (PSF) scores, whereas the Boston schools that administered the DIBELS did not. Buffalo schools did not report scores for the DIBELS Oral Reading Fluency and Retell Fluency (DORF), whereas Boston schools did.
    - The Boston school that used the MAP assessment provided only a percentile score for each time period. The Boston school that used Fountas & Pinnell provided a singular benchmark score.
    - Compared to Buffalo and Macon districts, the Boston school districts varied greatly in their reporting of benchmark scores. For both Fountas & Pinnell and DIBELS scores, a large share of Boston students were missing BOY, MOY, and EOY benchmark data.

Notes on other data provided by each district:

- **Race/ethnicity** was reported across all districts within the same general categories: Asian/Pacific Islander, Black/African-American, Hispanic/Latinx, White/Caucasian. Students who were reported as being "Multi-racial or other," "Mixed," "Two or more races," or "Multi-race not-Hispanic" were re-coded to the category of "Multi-racial or other."
- **English learner status** was reported as a yes/no variable from each district.
- **Gender** was reported across all districts as male or female.
- **Economic disadvantage** was constructed as a 0/1 variable. Two districts provided free and reduced-price lunch status, and one district provided an "economic disadvantage" indicator. One district did not provide any variables related to economic disadvantage, but all of its students received free lunch. Across districts, students who did not receive free or reduced-price lunch but were labelled as receiving McKinney-Vento Act services were included as part of the economically disadvantaged group.
- **Disciplinary actions** were recorded differently by all of the participating districts. All four districts provided in-school and out-of-school suspension records. Two districts provided any record of disciplinary action, which included warnings. Using these data, the evaluation team

created a 0/1 discipline flag to standardize disciplinary referrals across all participating districts: 0 indicated there were no records of any form of disciplinary action, and 1 indicated any record of disciplinary actions. Actions included warnings or suspensions.

- **Attendance** records were received from every district, though the amount of information reported varied greatly from district to district. The singular common variable across all districts was number of days a student was absent from school; however, a common denominator was not available.

## *Implementation/Fidelity Data Sources*

### *Tutor Observations*

The evaluation team conducted tutoring session observations during spring 2019. Observations were guided by an Experience Corps Observation Tool developed by the team using the Experience Corps session structure guidance (Appendix B). This tool was meant to capture information on:

- Session length
- Group size
- Tutoring location
- Tutor communications/collaborations with teachers
- Communications between tutor and student
- Skill-building session components: reading fluency, quality talk, tutor read-aloud
- Curriculums/materials used during the session
- Student engagement/distraction during the session

Program staff were asked to recommend two tutors from each school for the evaluation team to observe—one experienced tutor and one new tutor. The team attempted to observe at least two tutors in each participating school. Sometimes they were able to observe more than two tutors in a school (as time allowed); sometimes they were able to observe only one tutor (e.g., when tutors were absent).

The evaluation team observed a total of 67 Experience Corps tutoring sessions across the three local programs (Exhibit C-8 below). This included observations of 41 different tutors (in some cases, the same tutor was observed over multiple sessions).

**Exhibit C-8. Observations of Tutoring Sessions and Tutors**

| | Total | Macon | Buffalo | Boston |
|---|---|---|---|---|
| **Number of Sessions** | 67 | 31 | 18 | 18 |
| **Number of Tutors** | 41 | 17 | 11 | 13 |

### *Surveys*

In spring 2019, teachers and principals at study schools and Experience Corps tutors and program directors were invited to complete surveys to share their perspectives and experiences. Local Experience

Corps program staff provided the evaluation team with contact information for potential survey participants.

Across the four surveys, questions were designed to capture the most relevant information on a wide range of topics about implementation of the Experience Corps program model as well as anticipated outcomes. Experience Corps staff also provided input on the question types and content.

- There were two versions of the **teacher survey**, one for teachers in control classrooms and another for teachers in treatment classrooms. Teachers in treatment classrooms were asked about their communications with Experience Corps staff and tutors, the perceived effectiveness of Experience Corps, and tutoring session content/materials. Teachers in control classrooms were asked similar questions about non-Experience Corps literacy supports offered to their students. In the fall 2018, the teacher survey was administered in tandem with the DESSA-mini in which teachers rated their students' social-emotional learning.
- The **principal survey** covered the school's non-Experience Corps literacy and social-emotional learning supports, communications and relationships with Experience Corps staff, perceived program benefits, and the likelihood of the principal to recommend Experience Corps.
- The **tutor survey** asked about their experience and demographics, session structure and content, interactions with students, perceived benefits, and coaching/feedback. Tutor contact information was provided to the evaluation team by local program staff.
- The Experience Corps **program director survey** focused on tutor recruitment/background, trainings, tutoring delivery, and program costs. One survey was administered for each of the local programs in the study.

All four survey types were delivered on SurveyGizmo, an online software tool for survey programming and administration. Surveys were pre-populated with respondent information, such as name, program, school, etc. Most survey questions were multiple-choice or check-box style, though each survey also included a few open-ended questions.

Surveys were administered in May-June 2019 according to schedules determined to be most conveniently timed for each program and respondent type. Teachers, principals, program directors, and tutors were sent survey invitations by the evaluation team on the recommended date, followed by several reminders over the following weeks. Reminders to complete the survey came from both the Abt evaluation team and local program staff. The surveys remained open for four to six weeks, allowing participants ample time to complete them. The individualized survey links generated by SurveyGizmo allowed participants to stop and then return to their survey multiple times if desired.

The evaluation team responded to survey participants' questions throughout the administration period, directing questions to local program staff when needed. These local staff were instrumental in consistently encouraging teachers, principals, and tutors to complete their survey. The team provided local program staff with weekly updates on response rates; a summary is provided in Exhibit C-9.

**Exhibit C-9. Survey Respondents, by Local Program and Survey Type**

| Sample | # Surveys Administered | # Surveys Completed | Response Rate |
|---|---|---|---|
| **Teachers (Treatment and Control)** | | | |
| **Boston** | 68 | 34 | 50% |
| **Buffalo** | 74 | 49 | 66% |
| **Macon** | 97 | 79 | 81% |
| **Total** | **239** | **162** | **68%** |
| **Principals** | | | |
| **Boston** | 6 | 3 | 50% |
| **Buffalo** | 6 | 6 | 100% |
| **Macon** | 9 | 9 | 100% |
| **Total** | **21** | **18** | **86%** |
| **Tutors** | | | |
| **Boston** | 38 | 24 | 63% |
| **Buffalo** | 48 | 42 | 88% |
| **Macon** | 39 | 21 | 54% |
| **Total** | **125** | **87** | **70%** |
| **Program Directors** | | | |
| **Boston** | 1 | 1 | 100% |
| **Buffalo** | 1 | 1 | 100% |
| **Macon** | 1 | 1 | 100% |
| **Total** | **3** | **3** | **100%** |

### *Session Logs*

After each tutoring session, Experience Corps tutors record information about their session, including the students who attended and the duration (i.e., start time, end time). This information is entered into Salesforce by program staff and was shared with the evaluation team. The session logs provided valuable information about the dosage of Experience Corps sustained small group tutoring that students received.

## Analysis

The confirmatory (i.e., primary) outcome analysis was focused on the results of five of the six RAPID subtests: Phonological Awareness, Word Reading, Vocabulary Pairs, Following Directions, and Spelling. The evaluation team analyzed the effect of sustained small group tutoring on students' EOY scores by predicting their performance score from their classroom's assigned experimental condition (i.e., control or treatment), controlling for student pretest scores and student-level demographics (gender, grade, special education status, race/ethnicity, and English learner status). Students were nested within classrooms.

Even though the original evaluation plan called for a three-level model, the level 3 unit (schools) had at most an intraclass correlation of .035, suggesting that a third level would not add to the explanatory power of the model. As an exploratory measure, the evaluation team created a RAPID composite score

by averaging standardized versions[21] of students' performance scores across the subtests they were administered.

The primary analysis to test the effect of Experience Corps used a two-level HLM, with students nested within schools, to compare BOY and EOY RAPID scale scores. The HLM took the following form:

***Student Level:*** $$Y_{ij} = \beta_0 + \beta_{1j}(Y^*_{ij}) + \sum_{m=1}^{6} \beta_{2mj} X_{mij} + e_{ij}$$

***Classroom Level:*** $$\beta_{0j} = \gamma_{00} + \gamma_{01}(T_j) + \sum_{p=1}^{P-1} \gamma_{02.p} Block_{pj} + \mu_{0j}$$

In the student-level equation above,

$Y_{ij}$ is the posttest outcome for the $i^{th}$ student in the $j^{th}$ classroom,

$Y^*_{ij}$ is the pretest outcome for the $i^{th}$ student in the $j^{th}$ classroom,

$\beta_{1j}$ is the effect of pretest in class *j,*

$X_{mij}$ is the $m^{th}$ of six additional covariates for student *i* in classroom *j*:

- Baseline RSP score
- Grade
- Race/ethnicity
- Gender
- Special education status
- English learner status

$\beta_{2mj}$ is the effects of student covariates in classroom *j, and*

$e_{ij}$ is a residual error term for student *i* in classroom *j.*

In the class-level equation,

$\gamma_{00}$ is the mean intercept,

$\gamma_{01}$ is the treatment effect,

$\gamma_{02}$ is the effect of randomization block *p,*

and $\mu_{0j}$ is the random intercept term.

[21] To standardize performance scores for the creation of the composite, the evaluation team created z-scored versions of each student's individual subtest performance score by subtracting the mean across students and dividing by the standard deviation.

# Appendix D: Changes to the SIF Evaluation Plan

June 12, 2018

## Experience Corps

*Randomized Controlled Trial Design and Implications for Level of Evidence Rating*

In our August 2016 SIF Evaluation Plan (SEP), the AARP Foundation and its evaluator (Abt Associates) described plans for a randomized controlled trial (RCT) to take place in the 2018-19 school year. The impact evaluation design targeted strong evidence through use of a multi-site RCT with classroom-level assignment, blocked by school and grade level. We proposed a three-arm RCT, comparing the effectiveness of sustained small group tutoring alone and sustained small group tutoring plus literacy assistance (Blended model) to a comparison group which would not receive any Experience Corps services (Control). Based on recent internal discussions and past discussions with grantees, we have opted to move the study to a two-arm design, focusing on the sustained small group tutoring-only vs. Control comparison as our confirmatory contrast. The value-added of literacy assistance will still be investigated as an exploratory contrast, but this extension of the sustained small group tutoring model will not be randomly assigned as a condition.

Our motivation for this change was based on four factors:

1. Moving from a three-arm design to a two-arm design improves the statistical power of our study.
2. By allowing sites to implement literacy assistance where they are able, we can provide more flexibility to sites (i.e., they do not need to reserve one classroom for sustained small group tutoring, and they can implement literacy assistance where they have the capacity to do so).
3. It simplifies the random assignment process, since some schools have two classrooms per grade level and we no longer have to balance the roll-out of literacy assistance across blocks.
4. The direct contrast of sustained small group tutoring and sustained small group tutoring plus literacy assistance is not our primary research question, which instead focused on the effectiveness of the small group tutoring model. If the sustained small group tutoring strategy is found to be effective, the AARP Foundation can serve more students at lower cost. The study of literacy assistance – which was never intended to be tested as its own condition – does not have as important policy implications for the AARP Foundation relative to sustained small group tutoring.

All other procedures outlined in the SEP remain the same. We have secured the participation of the Boston, Buffalo, and Macon sites, which combined will include an anticipated 4 districts, 21 schools, 284 classrooms, and 3,244 students for the study. This will allow us to achieve a minimum detectable effect size of .183. In the 2018-19 school year, *after* students have been assigned to teachers using standard school practices, the Abt team will randomly assign classrooms within school and grade to either receive sustained small group tutoring or to serve in a Control condition and not receive any Experience Corps

services. Classrooms that are assigned to the treatment condition (i.e., sustained small group tutoring) may also implement the Blended model (sustained small group tutoring + literacy assistance), though only students who receive sustained small group tutoring in those classrooms will be included in the evaluation. Key student outcomes will be assessed using a standardized measure of reading performance (RAPID assessment), administered once near the beginning of the school year and again toward the end of the year. We will also track students' social-emotional development using the DESSA-Mini.

The move from a three-arm to two-arm design still maintains a Strong level of evidence, since the study remains a large, multisite RCT that can support causal conclusions. In fact, we believe this change will strengthen the study, since our statistical power has been improved with this change.

# Appendix E: Tutoring Observation Form

**Tutor Name**_____ **Teacher**_____

**Obs Initials**_____ **Start Time:**_____ **End Time:**_____

**Student Grades** *(circle all that apply)*: **K 1 2 3**
**School** _____ **Date**_____

**Observation Checklist: SIF EC Sustained Small Group Tutoring**

**QC Name:**_____

**QC Date:**_____

| | | | | | |
|---|---|---|---|---|---|
| **General Information and Session Preparation** | | | | | |
| Where did the tutoring session take place? | Classroom | Library | Rsrc Rm | Other | |
| How many students were present? | 1 Stus | 2 Stus | 3 Stus | 4 Stus | >= 5 Stus |
| Did the tutor check in with the classroom teacher before the session? | Yes | No | Don't Know | | |
| **Ongoing Communication** | | | | | |
| Does tutor make an early connection with the students? | Yes | No | | | |
| Does tutor establish behavior expectations? | Yes | No | | | |
| Does tutor communicate session or activity goals? | Yes | No | | | |
| **Reading Fluency (Skill Bldg –Part 1)** | | **Duration (mins)** | | | |
| Tutor focuses on a short section of text | Yes | No | | | |
| Students identify possible trouble words at any point | Yes | No | | | |
| Tutor/stus progress through 3-5 fluency reading games or activities | Yes | No | [____ # of games] | | |
| Tutor checks student fluency during or at the end of session | Yes | No | | | |
| **Review/Reflect & Quality Talk (Skill Bldg – Part 2)** | | **Duration (mins)** | | | |
| Student(s) reflect on reading performance /ability (pos or neg) | Yes | No | | | |
| Tutor reflects on student(s) effort | Yes | No | | | |
| Conversation strengthens comprehension, vocab, critical thinking | Yes | No | | | |
| **Tutor Read-Aloud (Skill Bldg – Part 3)** | | **Duration (mins)** | | | |
| Tutor states purpose for read-aloud | Yes | No | | | |
| Tutor models good reading of text (at least 2 sentences) | Yes | No | | | |
| **Session Characteristics** | | | | | |
| Did tutor use a curriculum / leveled reader at any point? | Yes | No | | | |
| If YES, identify the curriculum and/or source of materials? | Reading A to Z | Other | | | |
| What was the overall level of student engagement? | Engaged | Neutral | Disengaged | Mixed | |
| What was the level of external distraction? | Extensive | Moderate | Infrequent | None | |
| How frequently did the tutor provide behavior management support? | Extensive | Moderate | Infrequent | None | |

# Appendix F: Multi-level Model Results

**Exhibit F-1: Baseline Data, Confirmatory Analyses**

| | | Treatment Group (Experience Corps ) | | | Control Group | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Grades | N | Unadjusted Mean | SD | N | Mean | SD | Adjusted Difference* | Effect Size | *p* |
| **Phonological Awareness** | K | 126 | 203.40 | 114.12 | 73 | 199.64 | 115.90 | −12.95 | −.11 | .441 |
| **Word Reading** | 1-3 | 338 | 479.49 | 155.00 | 167 | 509.78 | 125.59 | −16.60 | −.11 | .066 |
| **Vocabulary Pairs** | K-3 | 464 | 453.96 | 107.12 | 239 | 450.12 | 109.44 | −2.50 | −.02 | .696 |
| **Following Directions** | 1-3 | 338 | 462.23 | 132.35 | 167 | 484.20 | 136.12 | −9.12 | −.07 | .377 |
| **Spelling** | 2-3 | 216 | 526.43 | 132.49 | 111 | 552.13 | 135.09 | −23.54 | −.18 | .061 |
| **Reading Success Profile (RSP)** | K-3 | 491 | 20.47 | 23.82 | 262 | 23.53 | 25.97 | −3.68 | −.15 | .016 |
| **DESSA-mini** | K-3 | 316 | 53.21 | 10.68 | 170 | 51.26 | 10.87 | 1.43 | .18 | .162 |

* The baseline model only uses the treatment indicator and the blocking variable as covariates.

**Exhibit F-2: Posttest Data, Confirmatory Analyses**

| | | Treatment Group (Experience Corps ) | | | Control Group | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Grades | N | Unadjusted Mean | SD | N | Mean | SD | Adjusted Difference* | Effect Size | *p* |
| **Phonological Awareness** | K | 126 | 437.52 | 179.19 | 73 | 418.47 | 195.07 | −40.56 | −.22 | .202 |
| **Word Reading** | 1-3 | 338 | 541.07 | 110.80 | 167 | 566.69 | 105.69 | −15.54 | −.14 | .108 |
| **Vocabulary Pairs** | K-3 | 464 | 494.74 | 123.77 | 239 | 488.76 | 109.81 | 1.43 | .01 | .872 |
| **Following Directions** | 1-3 | 338 | 482.18 | 139.36 | 167 | 498.83 | 142.00 | −2.68 | −.02 | .825 |
| **Spelling** | 2-3 | 216 | 599.19 | 136.24 | 111 | 613.91 | 108.89 | −4.53 | −.04 | .784 |
| **DESSA-mini** | K-3 | 316 | 55.38 | 10.83 | 170 | 54.04 | 12.80 | −0.20 | −.02 | .869 |

* Adjusted using the treatment indicator, pretest measure, baseline RSP score, grade, race/ethnicity, gender, special education status, and English learner status as covariates.

**Exhibit F-3: Exploratory Findings from Benchmark Data (OLS Regression Findings)**

| | Local Program | N | Experience Corps Estimate | *p* |
|---|---|---|---|---|
| **Star 360 Early Literacy** | Macon | 87 | −3.00 | .407 |
| **Star 360 Reading** | Macon | 55 | −0.88 | .754 |
| **DIBELS** | Buffalo | 309 | −8.00 | .261 |
| **DIBELS ORF-Accuracy** | Boston | 115 | −7.08 | .122 |

**Exhibit F-4: Multi-level Model Results: Phonological Awareness**

| | Descriptor | Estimate | Standard Error | DF | t Value | Pr > \|t\| |
|---|---|---|---|---|---|---|
| **Intercept** | | 411.69 | 108.64 | 38 | 3.79 | 0.0005 |
| **Condition** | | −40.5604 | 31.6095 | 130 | −1.28 | 0.2017 |
| **PAPerfScore_t1** | | 0.2004 | 0.1728 | 130 | 1.16 | 0.2483 |
| **RSP_t1** | | 2.4167 | 1.2464 | 130 | 1.94 | 0.0547 |
| **race_ethnicity** | **Asian** | −126.05 | 57.2261 | 130 | −2.20 | 0.0294 |
| **race_ethnicity** | **Black** | −127.02 | 43.5317 | 130 | −2.92 | 0.0042 |
| **race_ethnicity** | **Hispanic/Latinx** | −103.42 | 36.8805 | 130 | −2.80 | 0.0058 |
| **race_ethnicity** | **Multi-Racial or Other** | −116.26 | 81.0268 | 130 | −1.43 | 0.1537 |
| **race_ethnicity** | **White** | 0 | . | . | . | . |
| **gender** | | −8.4095 | 24.2304 | 130 | −0.35 | 0.7291 |
| **sped_status** | | −55.1653 | 53.7840 | 130 | −1.03 | 0.3069 |
| **English learner** | | 41.1591 | 34.8487 | 130 | 1.18 | 0.2397 |

**Exhibit F-5: Multi-level Model Results: Word Reading**

| | Descriptor | Estimate | Standard Error | DF | t Value | Pr > \|t\| |
|---|---|---|---|---|---|---|
| **Intercept** | | 378.02 | 42.2821 | 107 | 8.94 | <.0001 |
| **Condition** | | −16.1451 | 9.6887 | 329 | −1.67 | 0.0966 |
| **Grade_num** | **1** | −64.3127 | 52.0041 | 329 | −1.24 | 0.2171 |
| **Grade_num** | **2** | 121.03 | 67.3665 | 329 | 1.80 | 0.0733 |
| **Grade_num** | **3** | 0 | . | . | . | . |
| **WReadPerfScore_t1** | | 0.3254 | 0.04048 | 329 | 8.04 | <.0001 |
| **RSP_t1** | | 0.5702 | 0.1914 | 329 | 2.98 | 0.0031 |
| **race_ethnicity** | **Asian** | 25.0096 | 17.0346 | 329 | 1.47 | 0.1430 |
| **race_ethnicity** | **Black** | −9.1041 | 12.5260 | 329 | −0.73 | 0.4679 |
| **race_ethnicity** | **Hispanic/Latinx** | −0.9093 | 11.5747 | 329 | −0.08 | 0.9374 |
| **race_ethnicity** | **Multi-Racial or Other** | 9.7859 | 19.7040 | 329 | 0.50 | 0.6198 |
| **race_ethnicity** | **White** | 0 | . | . | . | . |
| **gender** | | −16.3330 | 7.0105 | 329 | −2.33 | 0.0204 |
| **sped_status** | | −8.6272 | 12.8516 | 329 | −0.67 | 0.5025 |
| **English learner** | | −8.4174 | 10.8831 | 329 | −0.77 | 0.4398 |

**Exhibit F-6: Multi-level Model Results: Vocabulary Pairs**

| Effect | Descriptor | Estimate | Standard Error | DF | t Value | Pr > \|t\| |
|---|---|---|---|---|---|---|
| **Intercept** | | 427.38 | 45.8333 | 146 | 9.32 | <.0001 |
| **Condition** | | 1.4342 | 8.9077 | 467 | 0.16 | 0.8722 |
| **Grade_num** | **0** | −140.49 | 64.7284 | 467 | −2.17 | 0.0305 |
| **Grade_num** | **1** | −108.76 | 55.4689 | 467 | −1.96 | 0.0505 |
| **Grade_num** | **2** | −109.73 | 74.3444 | 467 | −1.48 | 0.1406 |
| **Grade_num** | **3** | 0 | . | . | . | . |
| **VPPerfScore_t1** | | 0.3160 | 0.04894 | 467 | 6.46 | <.0001 |
| **RSP_t1** | | 0.8789 | 0.1940 | 467 | 4.53 | <.0001 |
| **race_ethnicity** | **Asian** | 15.1396 | 16.3230 | 467 | 0.93 | 0.3541 |
| **race_ethnicity** | **Black** | −12.8051 | 12.2706 | 467 | −1.04 | 0.2972 |
| **race_ethnicity** | **Hispanic/Latinx** | −1.7846 | 11.1003 | 467 | −0.16 | 0.8723 |
| **race_ethnicity** | **Multi-Racial or Other** | 38.8569 | 19.9782 | 467 | 1.94 | 0.0524 |
| **race_ethnicity** | **White** | 0 | . | . | . | . |
| **gender** | | −2.3179 | 6.7884 | 467 | −0.34 | 0.7329 |
| **sped_status** | | 7.0956 | 13.0533 | 467 | 0.54 | 0.5870 |
| **English learner** | | −20.1937 | 10.4243 | 467 | −1.94 | 0.0533 |

**Exhibit F-7: Multi-level Model Results: Following Directions**

| Effect | Descriptor | Estimate | Standard Error | DF | t Value | Pr > \|t\| |
|---|---|---|---|---|---|---|
| **Intercept** | | 303.20 | 56.4157 | 107 | 5.37 | <.0001 |
| **Condition** | | −2.6818 | 12.1365 | 329 | −0.22 | 0.8253 |
| **Grade_num** | **1** | −36.1607 | 64.9572 | 329 | −0.56 | 0.5781 |
| **Grade_num** | **2** | −48.2459 | 91.3876 | 329 | −0.53 | 0.5979 |
| **Grade_num** | **3** | 0 | . | . | . | . |
| **FDPerfScore_t1** | | 0.4707 | 0.05297 | 329 | 8.89 | <.0001 |
| **RSP_t1** | | 0.6902 | 0.2816 | 329 | 2.45 | 0.0148 |
| **race_ethnicity** | **Asian** | 5.9875 | 25.7218 | 329 | 0.23 | 0.8161 |
| **race_ethnicity** | **Black** | −22.9036 | 19.0921 | 329 | −1.20 | 0.2311 |
| **race_ethnicity** | **Hispanic/Latinx** | −3.8997 | 17.6151 | 329 | −0.22 | 0.8249 |
| **race_ethnicity** | **Multi-Racial or Other** | 1.5977 | 29.6908 | 329 | 0.05 | 0.9571 |
| **race_ethnicity** | **White** | 0 | . | . | . | . |
| **gender** | | 6.8305 | 10.6629 | 329 | 0.64 | 0.5222 |
| **sped_status** | | 2.6052 | 19.1721 | 329 | 0.14 | 0.8920 |
| **English learner** | | −29.5015 | 16.4137 | 329 | −1.80 | 0.0732 |

**Exhibit F-8: Multi-level Model Results: Spelling**

| Effect | Descriptor | Estimate | Standard Error | DF | t Value | Pr > \|t\| |
|---|---|---|---|---|---|---|
| **Intercept** | | 295.36 | 61.2318 | 72 | 4.82 | <.0001 |
| **Condition** | | −4.5250 | 16.4705 | 207 | −0.27 | 0.7838 |
| **Grade_num** | **2** | 49.6493 | 86.1802 | 207 | 0.58 | 0.5652 |
| **Grade_num** | **3** | 0 | . | . | . | . |
| **SPPerfScore_t1** | | 0.4918 | 0.07185 | 207 | 6.85 | <.0001 |
| **RSP_t1** | | 0.2298 | 0.3548 | 207 | 0.65 | 0.5179 |
| **race_ethnicity** | **Asian** | 63.8360 | 24.0080 | 207 | 2.66 | 0.0085 |
| **race_ethnicity** | **Black** | 16.5031 | 18.7850 | 207 | 0.88 | 0.3807 |
| **race_ethnicity** | **Hispanic/Latinx** | 31.0801 | 16.9537 | 207 | 1.83 | 0.0682 |
| **race_ethnicity** | **Multi-Racial or Other** | 52.1419 | 27.7380 | 207 | 1.88 | 0.0615 |
| **race_ethnicity** | **White** | 0 | . | . | . | . |
| **gender** | | −4.9398 | 10.2687 | 207 | −0.48 | 0.6310 |
| **sped_status** | | −27.6036 | 19.6927 | 207 | −1.40 | 0.1625 |
| **English learner** | | −8.3518 | 15.8563 | 207 | −0.53 | 0.5990 |

**Exhibit F-9: Exploratory Findings on Dosage and Program-Level Results**

| Effect | Descriptor | Estimate | Standard Error | DF | t Value | Pr > \|t\| |
|---|---|---|---|---|---|---|
| **Intercept** | | −0.1655 | 0.6707 | 74 | −0.25 | 0.8058 |
| **Condition** | | 0.06408 | 0.3099 | 293 | 0.21 | 0.8363 |
| **Grade_num** | **0** | −1.0381 | 0.8985 | 293 | −1.16 | 0.2489 |
| **Grade_num** | **1** | −0.6359 | 0.8398 | 293 | −0.76 | 0.4495 |
| **Grade_num** | **2** | 0.1597 | 0.3730 | 293 | 0.43 | 0.6688 |
| **Grade_num** | **3** | 0 | . | . | . | . |
| **average_rapid_t1** | | 0.3783 | 0.06926 | 293 | 5.46 | <.0001 |
| **RSP_t1** | | 0.005845 | 0.001925 | 293 | 3.04 | 0.0026 |
| **race_ethnicity** | **Asian** | 0.07433 | 0.1313 | 293 | 0.57 | 0.5718 |
| **race_ethnicity** | **Black** | −0.1410 | 0.09542 | 293 | −1.48 | 0.1405 |
| **race_ethnicity** | **Hispanic/Latinx** | −0.07237 | 0.08780 | 293 | −0.82 | 0.4105 |
| **race_ethnicity** | **Multi-Racial or Other** | 0.2697 | 0.1652 | 293 | 1.63 | 0.1036 |
| **race_ethnicity** | **White** | 0 | . | . | . | . |
| **gender** | | −0.02440 | 0.05688 | 293 | −0.43 | 0.6683 |
| **sped_status** | | −0.1126 | 0.1046 | 293 | −1.08 | 0.2824 |
| **Experience Corps Minutes** | | 0.000195 | 0.000094 | 293 | 2.07 | 0.0395 |
| **ELL** | | −0.1947 | 0.08503 | 293 | −2.29 | 0.0227 |
| **LocalProgram** | **Boston** | −2.2923 | 0.7764 | 293 | −2.95 | 0.0034 |
| **LocalProgram** | **Buffalo** | −0.2439 | 0.6383 | 293 | −0.38 | 0.7026 |
| **LocalProgram** | **Macon** | 0 | . | . | . | . |

# References

The Annie E. Casey Foundation. (2010). *Early warning! Why reading by the end of third grade matters*. Baltimore, MD: Leila Fiester. Retrieved from http://www.nmvoices.org/attachments/natl_kc_learn_to_read_full_rpt_5-10.pdf

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society Series B Methodological*, *57*(1), 289-300.

Carlson, M.C., Erickson, K.I., Kramer, A.F., Voss, M.W., Bolea, N., Mielke, M.,… Fried, L.P. (2009). Evidence for neurocognitive plasticity in at-risk older adults: The Experience Corps program. *Journals of Gerontology Series A Biological Sciences and Medical Sciences, 64*(12), 1275-1282.

Carlson, M.C., Kuo, J.H., Chuang, Y., Varma, V.R., Harris, G., Albert, M.S.,… Fried, L.P. (2015). Impact of the Baltimore Experience Corps trial on cortical and hippocampal volumes. *Alzheimer's & Dementia,* 1-9.

Cohen, J. (1988). *Statistical power analysis for the behavioral Sciences.* New York: Routledge Academic.

Crosby, J.W. (2011). Test Review: F.M. Gresham & S.N. Elliott Social Skills Improvement System Rating Scales. Minneapolis, MN: NCS Pearson, 2008. *Journal of Psychoeducational Assessment*, *29*(3), 292-296. doi:10.1177/0734282910385806

Experience Corps. (2014). *Experience Corps small group pilot: Implementation evaluation.* Washington, DC: AARP Foundation.

Foorman, B., Petscher, Y., & Schatschneider, C. (2019). Technical Manual for Lexia RAPID Assessment Version 5.0: Grades K-2. Concord, MA: Lexia Learning. Retrieved from http://www.lexialearningresources.com/RAPID/RAPID_TechnicalK2.pdf

Fried, L.P., Carlson, M.C., Freedman, M., Frick, K.D., Glass, T.A., Hill, J.,… Zeger, S. (2004). A social model for health promotion for an aging population: Initial evidence on the Experience Corps model. *Journal of Urban Health: Bulletin of the New York Academy of Medicine, 81*(1), 64-78.

Lee, Y.S., Morrow-Howell, N., Jonson-Reid, M., & McCrary, S. (2010). The effect of the Experience Corps® program on student reading outcomes. *Education and Urban Society, 44*(1), 97-118. doi:10.1177/0013124510381262

Lipsey, M.W., Puzio, K., Yun, C., Hebert, M.A., Steinka-Fry, K., Cole, M.W., Roberts, M., Anthony, K.S., & Busick, M.D. (2012). *Translating the statistical representation of the effects of education interventions into more readily interpretable forms* (NCSER 2013-3000). Washington, DC: National Center for Special Education Research, Institute of Education Sciences, U.S. Department of Education.

Porowski, A., de Mars, M., Kahn-Boesel, E., & Rodriguez, D.B. (2019). *Experience Corps social-emotional learning pilot evaluation.* Rockville, MD: Abt Associates.

RAND Education Assessment Finder (2018). *Devereux Student Strengths Assessment-Mini (DESSA-mini).* Retrieved from: https://www.rand.org/education-and-labor/projects/assessments/tool/2009/devereux-student-strengths-asessment-mini-dessa-mini.html

Rebok, G.W., Carlson, M.C., Glass, T.A., McGill, S., Hill, J., Wasik, B.A., Lalongo, N., Frick, K.D., Fried, L.P., & Rasmussen, M.D. (2004). Short-term impact of Experience Corps® participation on children and schools: Results from a pilot randomized trial. *Journal of Urban Health, 81*(1), 79-93.

Schachter, J.P. (2004, March). *Geographical mobility: 2002 to 2003. Population Characteristics* (Current Population Report P20-549). Washington, DC: U.S. Census Bureau. https://www.census.gov/prod/2004pubs/p20-549.pdf

Slavin, R.E, Lake, C., Chambers, B., Cheung, A., & Davis, S. (2011). Effective programs for struggling readers: A best evidence synthesis. *Education Research Review, 6,* 1-26.

Torgesen, J., Myers, D., Schirm, A., Stuart, E., Vartivarian, S., Mansfield, W., Stancavage, F., Durno, D., Javorsky, R., & Haan, C. (2006). *National assessment of Title I: Interim report. Volume II: Closing the reading gap, first year findings from a randomized trial of four reading interventions for striving readers* (NCEE 2006-4002). Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance.