Center for Social Development

GEORGE WARREN BROWN SCHOOL OF SOCIAL WORK

Evaluation of Experience Corps

Student Reading Outcomes

Nancy Morrow-Howell, Melissa Jonson-Reid, Stacey McCrary,
YungSoo Lee, Ed Spitznagel

Data collection services provided by Mathematica Policy Research
Emily Dwoyer, Kathy Sonnenfeld, Susan Sprachman

Funded by The Atlantic Philanthropies

January 2009

No. 09-01

Campus Box 1196 One Brookings Drive St. Louis, MO 63130-9906 • (314) 935.7433 • www.gwbweb.wustl.edu/csd

Table of Contents

The study team at Washington University used the consult of a team of advisors on various aspects of this project. The study team would like to acknowledge these advisors, while retaining full responsibility for the contents of this report.

ADVISORY COMMITTEE MEMBERS

Mark Dynarski	Mathematica Policy Research Center for Improving Research Evidence Vice President and Director
Robert E. Eckardt	Cleveland Foundation Senior Vice President of Programs and Evaluation
Don Grantt	United States Administration on Aging Program Specialist, Center for Planning and Policy Development
Jean B. Grossman	Public/Private Ventures Senior Vice-President of Research
Margaret Mark	Margaret Mark Strategic Insight President
Susan Moses	Harvard School of Public Health Center for Health Communication Deputy Director
Guitele Nicoleau	Academy for Educational Development Chief of Party, USAID: Basic Education Program Senegal, West Africa
David Reuben	UCLA School of Medicine Archstone Foundation Professor of Medicine Chief, Division of Geriatrics Director, Multicampus Program in Geriatric Medicine and Gerontology
Steven P. Wallace	UCLA School of Public Health, Professor UCLA Center for Health Policy Research, Associate Director
Terrie Fox Wetle	Brown University Associate Dean of Medicine for Public Health and Public Policy Professor, Medical sciences

Executive Summary

Evaluation of Experience Corps: Student Reading Outcomes

The Experience Corps (EC) program brings older adults aged 55+ into public elementary schools to tutor and mentor children who are at risk of academic failure. The EC program began in 1995 in five cities and has grown to include 23 sites. Currently, there are nearly 2,000 EC tutors serving approximately 20,000 students. Older adults are recruited to serve in this program and receive training to prepare them for their service assignments, focused on literacy and relationship-building. Each Experience Corps volunteer, or "member," is assigned as part of a team to a local elementary school participating in the program. At the beginning of the school year, teachers refer low-achieving students to the program; and EC members begin regular tutoring with the children.

In 2006, researchers at the Center for Social Development at Washington University's Brown School of Social Work were awarded a grant from The Atlantic Philanthropies to evaluate the effects of the Experience Corps program on student reading outcomes. Mathematica Policy Research, Inc. (MPR) provided data collection services.

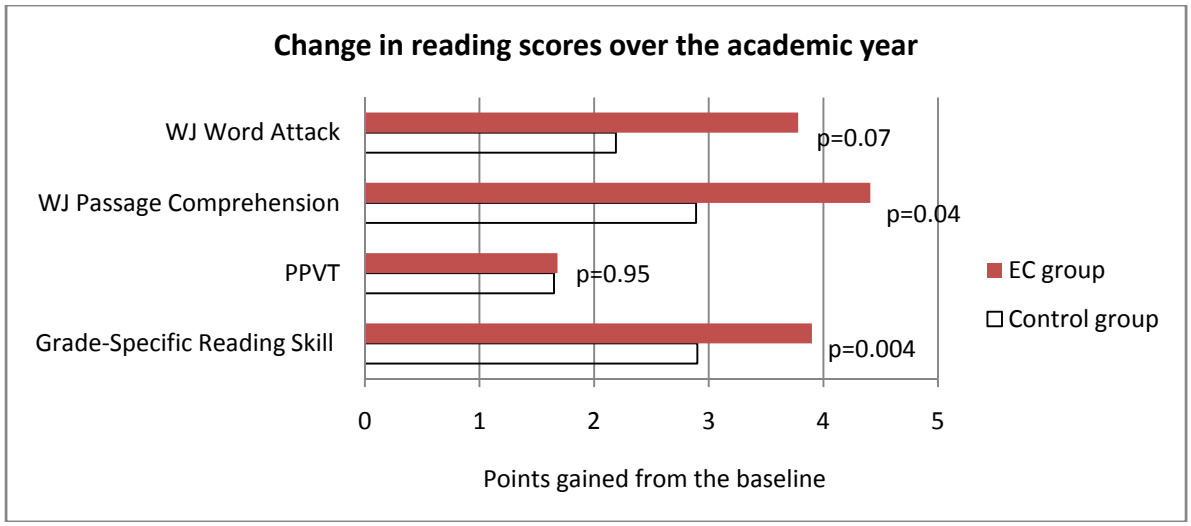
Twenty-three schools in Boston, New York City, and Port Arthur, Texas, participated in the study. A two group, pre-post test design with random assignment was used to assess the effects of the EC program. At the beginning of the school year, teachers referred all students who needed assistance with reading. Students were randomly assigned to the EC program, as there were not enough tutors to serve all of the referred students. Over 1,000 students were referred. Parental consent was obtained on 81% of the referred students, and 883 students were pretested. At posttest, 825 students were reassessed. The EC program tutored 430 of these students, and 451 were in the control group. There were 332 1st, 304 2nd, and 186 3rd graders; 420 males and 402 females in the final dataset.

Data for the study came from three sources: interviews with the students; assessments completed by teachers; and school records. MPR interviewers assessed reading ability at the beginning and end of the school year in face-to-face interviews with the students. Standardized reading tests were used: the Woodcock Johnson word attack subscale (WJ-WA), the Woodcock Johnson passage comprehension subscale (WJ-PC), and the Peabody Picture Vocabulary test (PPVT-III). These widely used measures were chosen because they were not specific to any one of the tutoring curricula used in the participating EC programs, but there was some correspondence between skills assessed by the standardized measures and aspects of the various programs' curricula. At the beginning and end of the academic year, teachers completed assessments of grade-specific reading skills and classroom behavior. At the end of the year, school records were abstracted to ascertain demographics and other student characteristics, and tutors rated the quality of their relationships with the EC students as well as provided their perceptions of student progress.

Analysis of pretest data showed that the EC students and control groups were equivalent on all measured characteristics. Students referred to the EC program were very poor readers and were clearly in need of assistance. From the scores on the WJ-PC measure, we can conclude that half of the students referred to EC perform as low as or lower than 84% of the students their age nationwide, and 12% score worse than 97% of the population.

The EC program succeeded in delivering the intervention to a large number of the students identified for the program. About half of the EC students received 30 to 49 sessions, and the mean number of sessions was 45. Three-quarters of the students received over 35 sessions, which represents about one session a week throughout the program period.

The students in the EC program made statistically greater gain over the academic year on passage comprehension and on assessments of grade-specific reading skills made by the teachers ($p < .05$); and the group difference on word attack was marginally significant ($p < .07$). Gain scores of the experimental and control group are displayed in the following chart.



In general, the effects of the program were consistent across subgroups of students. That is, the program impact was the same no matter what the gender, ethnicity, grade, classroom behavior, or English proficiency of the student. However, it is important to note that special education students, operationalized as those with IEPs in the student record, did not benefit from the EC program as much as non-special education students in regards to reading comprehension. EC programming with special education students needs to be reconsidered in light of this finding.

When including only the EC students who received at least 35 sessions, a criterion that we chose to indicate that the students received the intervention as intended, the effects were stronger. The effect sizes associated with the improvement in reading outcomes were .13 to .17.

Teachers overwhelmingly rated the EC program as beneficial to students, and they found that it had no or low burden to them. Tutors perceived that the EC program had a positive impact on students, and their relationships with students were good. Further, tutor relationship was related to reading outcomes, with better relationships associated with better outcomes.

In sum, these findings indicate that the EC program had statistically significant and substantively important effects on reading outcomes.

Evaluation of Experience Corps: Student Reading Outcomes

Researchers at the Center for Social Development at Washington University's Brown School of Social Work (WUSTL) were awarded a grant from The Atlantic Philanthropies to assess the effects of the Experience Corps (EC) program on both the students receiving the service and the older adults providing the service. This report includes the results of the research on the reading outcomes experienced by students participating in the EC program.

The Experience Corps (EC) program brings older adults aged 55+ into public elementary schools to tutor and mentor children who are at risk of academic failure. The EC program began in 1995 in five cities and has grown to include 23 sites. Currently, there are nearly 2,000 EC tutors serving approximately 20,000 students. Older adults are recruited to serve in this program and receive training to prepare them for their service assignments, focused on literacy and relationship-building. Each Experience Corps volunteer, or "member," is assigned as part of a team to a local elementary school participating in the program. At the beginning of the school year, teachers refer low-achieving students to the program; and EC members begin regular tutoring with the children.

Older adults are recruited to serve in this program via written advertisement and word-of-mouth. The volunteers are screened, interviewed, and receive training focused on literacy and relationship building. They are then assigned to local elementary schools participating in the program. At the beginning of the school year, teachers refer low reading students to the program, and EC members begin regular sessions with the children. They work with the students throughout the academic year. The large majority of members provide one-on-one tutoring, and most work about 15 hours per week. Across all program sites in the country, over two-thirds of the members receive a small stipend for this high-commitment role.

EC Programs in the Evaluation

This evaluation focused on the EC program in three cities: Boston, New York, and Port Arthur, Texas. These cities were chosen for several reasons: 1) they were long-running and established programs; 2) the research team could rely on stable administration and well-developed relationships with the schools; 3) these cities had programs large enough to yield the desired sample; and 4) the school districts in these cities gave approval for the research to be completed. Other EC programs across the country were eager to participate, but school district personnel were not willing to approve the research or not willing to allow randomization of the students. Also, some program sites were not large enough to supply enough students for the desired sample size, and others were too new to ensure stable relationships with the schools.

All of the EC schools in Port Arthur (eight schools) participated in the research. At the time of evaluation planning, there were 10 EC schools in Boston who were committed to participating in the EC program in the 2006-07 school year (during the planning stage, it was not yet certain if four additional schools would be participating in the EC program so they were not approached for the study). One out of the ten schools was excluded from the evaluation because the main mode of intervention with the students was not one-to-one tutoring. The remaining nine schools in Boston

participated in the evaluation. In New York, there were 16 EC schools, and 6 participated in the research. We selected only a sample of the New York schools to meet sample needs and represent the entire EC program in New York. The selected schools were administratively stable. They were geographically dispersed throughout the district and adequately represented all the schools that host the EC program according to overall school characteristics. For example, the average number of students at the selected schools was 520 and the average number of students at the non-selected EC schools was 538. Also, more than 90% of the students at selected schools were on free or reduced lunch (indicating low income families), and this is true in most of the New York schools participating in EC. The range of teacher/student ratio in schools involved in the study was 1:9-1:14 and it was 1:11-1:17 in the non-study schools.

The program differs in several ways between these cities. The size of each program varies, from about 50 volunteers in Port Arthur to about 150 volunteers in New York to over 300 in Boston. Further, EC members participate at different levels. In New York and Port Arthur, all members serve 15 hours a week, while members in Boston can participate at various levels. All EC members in New York and Port Arthur are stipended while some members in Boston are un-stipended. In regard to the work with the students, the curricula used in the tutoring sessions are different. All three programs serve 1st through 3rd grade, although Boston also serves 4th and 5th graders. All cities serve kindergartners, but with various emphases. In this evaluation, we focused on 1st through 3rd grades to achieve adequate subsample size by grade. Details of the Boston, New York, and Port Arthur EC programs are presented in Appendix A.

In regard to the essential elements of the EC program, there are similarities among the three sites. Across all three cities, the EC intervention is a one-to-one pull-out program—meaning the tutors work individually with children, most commonly in space outside of the classroom, but sometimes in a more private place in the classroom. Teachers refer students in need of reading assistance. The tutors use a structured curriculum and materials provided by the EC program. The EC members are generally recruited and screened in the same way. EC program coordinators in all three cities take applications, conduct interviews, check references and require a criminal background check. The program coordinators provide comprehensive training and on-going supervision of the tutors. There are regular support/training meetings with the EC staff and members, and EC members receive a performance evaluation. In all three programs, EC staff members provide coordination between the EC tutors and the classroom teachers.

Although the study was completed in three cities, these programs represent one-to-one tutoring activities with 1st through 3rd grades in EC programs across the country. The core model of the EC program nationally remains one-to-one tutoring, with 88% of volunteers across the country self-reporting that they perform that function, and the focus of intervention remains younger students in elementary school. The three sites participating in the study, as well as the other programs around the country, generally follow the key elements of a successful reading program outlined by Barbara Wasik (1998): a designated coordinator who knows about reading instruction; the presence of structure in the tutoring sessions; training of the tutors; and coordination between the volunteer program and classroom instruction. Although specific curricula differ across cities, the tutors are trained and supported in using a structured curriculum.

Study Methodology

The study of reading outcomes spanned two academic years. In 2006-07, the focus was on Boston and New York. In 2007-2008, Port Arthur was added (we were not able to accomplish all of the administrative permissions necessary to complete three program sites during the first year). The researchers from WUSTL worked with EC staff to develop a feasible and acceptable research strategy. They met with school district administrators and school principals to obtain permissions. WUSTL contracted with Mathematica Policy Research (MPR) in Princeton to receive the teachers' referrals, randomize referrals into the Experience Corps program or the control condition, conduct the interviews, and enter the data. WUSTL researchers then analyzed the data and produced final reports. All procedures were approved by the IRB at Washington University (E05-133).

Design

A two group, pre-post test design with random assignment was used to assess the effects of the EC program. At the beginning of the school year, teachers were asked to refer all students who needed assistance with reading and not constrain the list to match the capacity of the program. Thus, more students were referred than could be served. The names of the referred students were sent to MPR, who sent letters to parents, seeking written permission for the student to participate in the study. MPR applied a lottery system to the referred names to determine which students would be in the EC program (see Appendix B for randomization summary). The selected student names were sent to EC program coordinators to assign tutors and begin tutoring sessions. All study participants were pretested as early in the semester as possible. Pretesting occurred from mid-September to end of November. By the end of October, MPR had completed pretesting on 72% of the sample, in all three cities. We attempted to posttest all students beginning one month before the end of the school year, even if they had moved within the district during the academic year.

Sample

As seen in Table 1 on the next page, 1,100 students were referred by teachers. Parental consent was obtained on 81% of the referred students, and 883 students were pretested. At posttest, 825 students were located and reassessed. The EC program tutored 434 of these students, and 454 were in the control group. For data analysis, several observations were dropped due to missing birthdates or extensive missing data. This resulted in a final pretest sample size of 881 and posttest sample size of 822. There were 332 first graders, 304 second graders, and 186 third graders; 420 males and 402 females in the final dataset.

Data collection

Data on student demographics, reading, and related variables came from three sources: interviews with students; surveys completed by teachers; and school records. MPR staff assessed reading ability at the beginning of the school year. Students were taken from the classroom at times approved by the teachers and completed 30-minute face-to-face interviews. As recognition for participating, students selected school supplies at the end of both the pretest and posttest interviews.

MPR distributed surveys to teachers at the beginning and end of the school year. They received $15 for each survey completed and returned to MPR. Overall, the teachers provided information on 84% of the students, yet one school in Boston and two in Port Arthur had teacher participation rates less than 50%, despite on-going efforts by MPR to increase this response rate.

At the end of the academic year, school personnel abstracted school records, capturing student demographics and school behavior.

Table 1. Experience Corps Recruitment and Assessment Summary

	Total referrals	Consented		Student Assessments		Teacher Assessments	
	#	#	%	#	%	#	%
Boston							
Boston School #1	81	69	85%	69	100%	47	68%
Boston School #2	49	40	82%	40	100%	35	88%
Boston School #3	48	41	85%	41	100%	27	66%
Boston School #4	64	47	73%	47	100%	47	100%
Boston School #5	21	16	76%	16	100%	16	100%
Boston School #6	33	29	88%	29	100%	27	93%
Boston School #7	29	24	83%	24	100%	24	100%
Boston School #8	44	36	82%	36	100%	25	69%
Boston School #9	73	45	62%	45	100%	20	44%
Total:	442	347	79%	347	100%	268	77%
New York							
New York School #1	58	52	90%	52	100%	51	98%
New York School #2	56	41	73%	41	100%	31	76%
New York School #3	65	58	89%	58	100%	52	90%
New York School #4	64	46	72%	45	98%	46	100%
New York School #5	48	42	88%	42	100%	42	100%
New York School #6	63	54	86%	54	100%	46	85%
Total:	354	293	83%	292	100%	268	91%
Port Arthur							
Port Arthur School #1	27	24	89%	24	100%	24	100%
Port Arthur School #2	23	18	78%	18	100%	10	56%
Port Arthur School #3	37	33	89%	33	100%	25	76%
Port Arthur School #4	20	16	80%	16	100%	6	38%
Port Arthur School #5	75	64	85%	63	98%	56	88%
Port Arthur School #6	53	34	64%	32	94%	31	91%
Port Arthur School #7	41	33	80%	33	100%	28	85%
Port Arthur School #8	28	26	93%	25	96%	25	96%
Total:	304	248	82%	244	98%	205	83%
Total:	**1100**	**888**	**81%**	**883**	**99%**	**741**	**84%**

Measures

Standardized reading tests were used to capture student reading ability: the Woodcock Johnson word attack subscale (WJ-WA), Woodcock Johnson passage comprehension subscale (WJ-PC), and the Peabody Picture Vocabulary test (PPVT-III). These measures were chosen because they were

not specific to any one of the tutoring curricula used in the participating EC programs, but there was some correspondence between skills assessed by the standardized measures and aspects of the various programs' curriculum. Also, these measures are widely used in educational research. Finally, teacher assessment of grade-specific reading skills were also collected.

Woodcock Johnson III Tests of Achievement. The Woodcock Johnson III - Tests of Achievement (WJ III ACH) includes tests for written language, oral language, and academic knowledge (Woodcock, McGrew, & Mather, 2001; Gunn, B., Biglan, Smolkowski, & Ary, 2000). It is designed to measure intellectual abilities and academic achievement. There are two forms that consist of 22 tests subdivided into the standard and extended battery tests. Scoring is completed during testing to determine basal and ceiling levels. The reliability scores for the WJ III meet or exceed standards. Concurrent, construct, and criterion validity are indicated to be strong.

Two subtests of the full measure were used in this study. The WJ-WA sub-test assesses the student's phonemic awareness skills. Students were asked to read a list of nonsense words, such as "zoop" or "thrept." The WJ-PC sub-test assesses the student's overall skill at understanding text. Students silently read a short passage and then fill in the missing word. We chose these measures because they aligned with the curriculum of the New York program, which emphasized phonetics and comprehension, and the curriculum of the Boston program, which emphasized comprehension.

Peabody Picture Vocabulary Test (PPVT-III). The PPVT-III was chosen because it had been used in a previous study of EC and had picked up statistically significant changes in student ability (Rebok et al., 2004). The PPVT-III measures receptive or hearing vocabulary for Standard American English and estimates verbal ability. It is age-normed for 2.5 years to 90+ year-old people. It can be used to test preschool children's vocabulary acquisition, screen for giftedness and mental retardation, measure English language proficiency in individuals for whom English is not a primary language, test persons who have moderate visual disabilities, and in research studies. The administration time is 10-15 minutes. Test-retest reliability, internal consistency reliability, and criterion-related validity are all good (Lloyd & Dunn, 1997).

Grade-specific reading skills. This measure was developed for the purpose of this evaluation (by Dr. Melissa Jonson-Reid in consultation with Frank Pajares) and was completed by the teachers. It was a modification of a measure developed to assess self-efficacy of young readers (Pajares, 2002; Chapman & Tunmer, 2003). The list of 10 skills was grade specific. The task-specific questions were drawn from various curricula standards and reviewed by reading consultants at MPR. Because students referred for tutoring are likely to be behind grade level, tasks from prior grades were asked. For example, first grade teachers were asked about skills like sounding out letters while second grade teachers were asked about sounding out a word. Each skill was assessed on a four-point scale and summed to a total score reflecting the teacher's assessment of reading ability. Inter-item correlation was .90.

Classroom behavior. In addition to the assessment of reading skills noted above, a modified version of the Sutter-Eyberg Student Behavior Inventory-Revised (SESBI-R) was included in the teacher survey (Eyberg & Pincus, 1999). The SESBI-R is a brief teacher rating scale designed to measure disruptive behavior problems (ODD, ADHD, CD) in children and adolescents between the ages of 2 and 16 to determine if treatment is needed for behavior problems. The scale consists of 38 items, is completed by teachers, and is useful in the assessment of disruptive behaviors in the school

setting. The SES-BI was modified for this survey to include five additional questions, focusing on positive student behavior. Additionally, the scale was changed from a seven-point to a five-point scale: never, rarely, sometimes, often, always.

Data in school records. From the school records, we abstracted gender, date of birth, and racial/ethnic group. Further, information on attendance, free lunch, Individual Educational Plan (IEP) and Limited English Proficiency (LEP) status was obtained.

Sample attrition

Fifty-nine students – about 7% of the pretest sample – dropped out of the study. This attrition was equally distributed across the EC and control groups. Attrition from the EC and control groups was equal (30 EC students and 29 control students). Those who completed posttest did not differ from those who dropped out in terms of major demographic variables.

MPR provided a report at the end of each data collection period (Appendix C). Reports outlined sampling, randomization, consent, assessment procedures and completed interviews. This was done for each city, for both student and teacher assessments.

Appendix D presents a flow chart of sample participation.

Data analysis

Missing data. Missing data stemmed from two sources: student attrition where the entire posttest was missing and completed interviews where certain variables were missing. Missing data from both sources were imputed using a Markov Chain Monte Carlo (MCMC) multiple imputation method. Missing data were imputed separately for treatment and control groups. Five imputed datasets were created, and estimates reported throughout this study are those combined from the five imputed data sets.

Parameter estimation. The impacts of the EC program were estimated by comparing posttest scores for the EC and control groups which were adjusted for pretest scores and other covariates such as gender, ethnicity, grade, site, and classroom behavior. The adjusted posttest scores are tested for statistical difference and used to calculate effect size. Effect sizes were calculated using Hedge's G. To test the differential impact on subgroups of students, we added interaction terms to the full model.

Clustering effects. The data used in the current study have a hierarchical structure (e.g., students are clustered within classrooms, classrooms are clustered within schools). In these clustered data, outcomes of individuals within a same cluster are likely to be correlated, and a failure to incorporate within-cluster correlations into the analytic model leads to incorrect standard errors and p-values (Ballinger, 2004; Peters et al., 2003). Based on this notion, estimates and corresponding p-values are adjusted by the Generalized Estimating Equation (GEE) method. Also, to facilitate interpretations of the results, we additionally report effect sizes for all outcomes.

In the analysis, standardized scores were used for WJ-WA, WJ-PC, and PPVT. We also employed the weights provided by MPR to account for the specific randomization procedures employed.

Appendix E contains more details about the analytic approaches described above.

Findings

Table 2 presents a description of the sample at the baseline.[1] Randomization of the students into the EC program and the control group was effective in creating two equal groups in terms of main demographics and other variables such as school absences and classroom behavior. Also, reading abilities of the two groups at baseline were equivalent (none of the group differences were statistically significant). Thus, it is more likely that any differences in reading abilities at the end of the academic year were due to participation in the EC program.

Reading scores of the students referred to the EC program were very low. For example, on the Woodcock Johnson passage comprehension, 92% were below the nation-wide mean, with 50% being one standard deviation below the nation-wide mean, and 12% being two standard deviations below. Similarly, 62% of the students scored one standard deviation below the nation-wide mean on PPVT-III and 20% were two standard deviations below this mean. These findings indicate that the children being referred to EC are being correctly identified for the program and are in need of reading assistance. It is notable that one-quarter of the students referred to the program have English as their second language. Also, 14% are special education students, as they have IEPs in the student records. These attributes further signal the need for literacy support.

[1] This table is based on the raw data, where missing data are not imputed; thus sample size is different among the variables. In Appendix D, the sample description is presented on the imputed data.

Table 2. Sample description at pretest

	Total	EC	Control	Difference Test
Reading Outcomes				
WJ Word Attack (Standardized Score)	91.89 (20.45) N=881	91.89 (20.31) N=430	91.89 (20.61) N=451	t=0.00, p=1.00
WJ Passage Comprehension (Standardized Score)	84.41 (13.72) N=881	83.99 (13.59) N=430	84.84 (13.85) N=451	t=-0.92, p=0.36
PPVT (Standardized Score)	80.95 (13.98) N=881	80.87 (14.30) N=430	81.03 (13.65) N=451	t=-0.17, p=0.87
Grade-specific reading skills	2.38 (0.58) N=734	2.37 (0.56) N=359	2.39 (0.60) N=375	t=-0.58, p=0.56
Demographics				
Gender				
Male	451 (51%)	209 (49%)	242 (54%)	χ^2=2.25, df=1,
Female	430 (49%)	221 (51%)	209 (46%)	p=0.13
Race				
African American	473 (58%)	238 (60%)	235 (56%)	χ^2=2.38, df=2,
Hispanic Origin	299 (36%)	135 (34%)	164 (39%)	p=0.30
Others	47 (6%)	25 (6%)	22 (5%)	
Grade				
1st grade	363 (41%)	180 (42%)	183 (40%)	χ^2=2.52, df=2,
2nd grade	318 (36%)	162 (38%)	156 (35%)	p=0.28
3rd grade	200 (23%)	88 (20%)	112 (25%)	
Age	7.09 (1.11) N=881	7.09 (1.11) N=430	7.10(1.12) N=451	t=-0.15, p=0.88
School Events				
Free lunch				
Yes	766 (94%)	370 (93%)	396 (95%)	χ^2=0.85, df=1,
No	49 (6%)	27 (7%)	22 (5%)	p=0.36
IEP (Individualized Education Plan)				χ^2=0.20, df=1,
Yes	112 (14%)	53 (14%)	59 (15%)	p=0.65
No	665 (86%)	330 (86%)	335 (85%)	
LEP (Limited English Proficiency)				χ^2=1.15, df=1,
Yes	189 (24%)	87 (22%)	102 (25%)	p=0.28
No	604 (76%)	305 (78%)	299 (75%)	
Student Behaviors				
Classroom Behavior	3.56 (0.77) N=735	3.54 (0.77) N=360	3.58 (0.77) N=375	t=-0.72, p=0.47

Table 3 presents the pretest and posttest scores on the reading measures as well as the difference between the two scores, indicating reading gains over the academic year. Statistical tests indicate that both groups made positive gains on the reading scores. For example, the treatment group (EC) gained 3.78 points on the WJ word attack measure and the control group gained 2.47 points; both gains are statistically different from zero (a zero score meaning no improvement).

Table 3. Pretest and posttest reading scores and gains over the academic year

Outcome Variable	Group	Pre	Post	Gain Score: Difference between Pre and Post
WJ word attack	Treatment (N=430)	91.89 (20.31)	95.67 (15.94)	3.78*** [0.87]
	Control (N=451)	91.89 (20.61)	94.36 (16.55)	2.47** [0.79]
WJ passage comprehension	Treatment (N=430)	83.99 (13.59)	88.40 (11.88)	4.41*** [0.58]
	Control (N=451)	84.84 (13.85)	87.30 (12.18)	2.46*** [0.66]
PPVT	Treatment (N=430)	80.87 (14.30)	82.55 (12.93)	1.68*** [0.41]
	Control (N=451)	81.03 (13.65)	82.75 (12.26)	1.72*** [0.43]
Grade-specific reading skills	Treatment (N=430)	2.36 (0.57)	2.75 (0.60)	0.39*** [0.03]
	Control (N=451)	2.38 (0.62)	2.66 (0.66)	0.28*** [0.02]

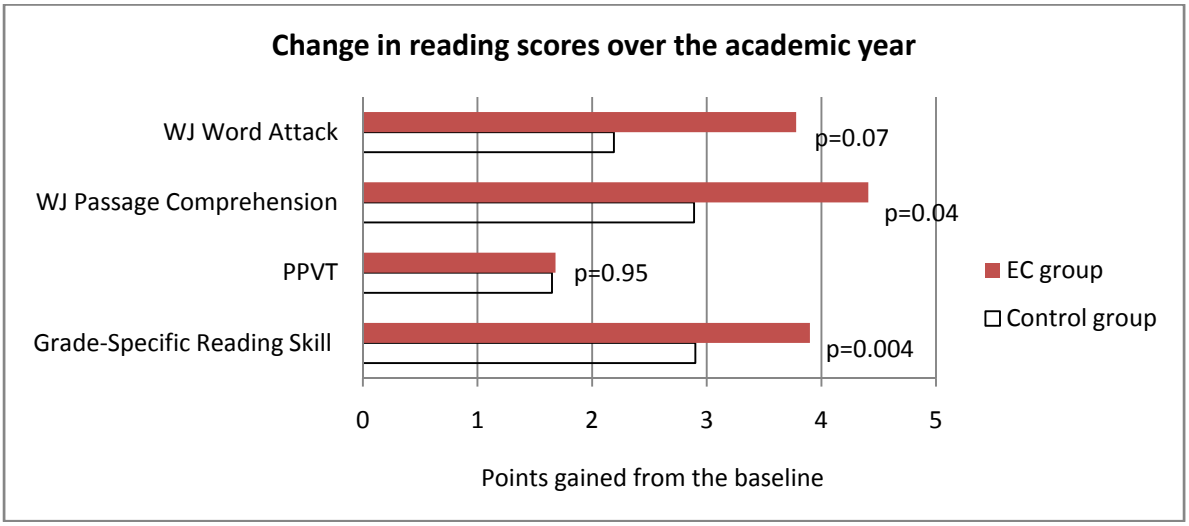
* p<.05 ** p<.01 *** p<.001
Numbers in parentheses are standard deviations; numbers in box brackets are standard errors

Table 4 presents findings on statistical tests of the differences between the gains experienced by the EC students and controls. The posttest scores were corrected for pretest scores as well as other covariates, including gender, ethnicity, grade, program site, classroom behavior, IEP, and LEP. On the WJ-passage comprehension measure and grade-specific reading skills, the changes made by EC students were statistically more positive than the changes made by control students ($p<.05$). Effect sizes associated with these gains are .13 and .16, respectively. An effect size of .16 indicates that the average gain of the EC students exceeded the gain of 56.4% of the control students. The group difference on word attack was marginally significant ($p < .07$), with an associated effect size of .10. The regression models yielding the adjusted posttest scores are included in Appendix F.

Table 4. Adjusted posttest reading scores and tests of significance

Outcome Variable	Treatment Adjusted posttest mean (N=430)	Control Adjusted posttest mean (N=451)	**Program Impact**	**Effect Size**
WJ word attack	95.79 [0.63]	94.20 [0.70]	**1.59†** **[0.88]** **t=1.80, p=0.07**	**0.10**
WJ passage comprehension	88.69 [0.45]	87.17 [0.57]	**1.52*** **[0.73]** **t=2.10, p=0.04**	**0.13**
PPVT	82.72 [0.39]	82.70 [0.35]	**0.03** **[0.51]** **t=0.06, p=0.95**	**0.002**
Grade-specific reading	2.76 [0.03]	2.66 [0.03]	**0.10**** **[0.03]** **t=3.02, p=0.004**	**0.16**

Numbers in brackets are standard errors.



The graph shows the gains made by EC students compared to controls. The bars for the EC group represent the difference between pretest and posttest scores and are the actual gain scores. The bars for the control group illustrate the gain for controls if matched for covariates. That is, the difference between the EC and control group gains is the estimated difference after adjustments for covariates.

Differential impacts of the EC program

To further specify the impact of EC, we explored whether some groups of EC students benefited more than others. We tested whether gender, grade, ethnicity, site, classroom behavior, being in special education (having an IEP), or having limited English proficiency (LEP) moderated program effects. Table 5 presents the parameter estimates and significance testing on the interaction terms which were used to test for moderating effects.

Table 5. Interaction effects between EC participation and other covariates

Interactions	WJ Word Attack	WJ Passage Comprehension	PPVT	Grade-specific reading skills
EC×female	0.46 [1.93] p=.81	-0.56 [1.37] p=.68	-0.25 [1.03] p=.81	-0.02 [0.09] p=.81
EC×1st grade	3.61 [2.17]	1.00 [1.71]	0.10 [1.22]	0.07 [0.07]
EC×3rd grade	-0.72 [2.22]	-0.13 [1.64]	-0.40 [1.21]	0.12 [0.10]
	F=2.22, df=2, p=.11	F=0.20, df=2, p=.82	F=0.08, df=2, p=.92	F=0.86, df=2, p=.43
EC×Boston	**-5.44 [2.17]**	-0.34 [1.94]	-2.31 [1.08]	-0.07 [0.09]
EC×Port Arthur	**-4.34 [2.33]**	-1.34 [1.97]	-1.78 [1.39]	-0.11 [0.09]
	F=3.28*, df=2, p=.04	F=0.31, df=2, p=.73	F=2.36, df=2, p=.10	F=0.80, df=2, p=.45
EC×Hispanic	2.35 [1.86]	0.98 [1.48]	-2.06 [1.03]	0.02 [0.09]
EC×other race	-4.58 [4.02]	-2.46 [2.76]	-1.95 [2.15]	-0.02 [0.19]
	F=1.65, df=2, p=.20	F=0.74, df=2, p=.48	F=1.94, df=2, p=.15	F=0.03, df=2, p=.97
EC×classroom behavior	1.33 [1.29] p=.31	1.33 [1.10] p=.23	-0.45 [0.62] p=.46	-0.01 [0.04] p=.77
EC×IEP	-1.51 [2.79] p=.59	**-4.75* [2.17] p=.03**	-0.03 [1.50] p=.98	0.06 [0.12] p=.61
EC×LEP	0.71 [1.90] p=.71	0.49 [1.58] p=.76	-1.10 [1.19] p=.36	0.001 [0.08] p=.99

Notes: Each posttest reading score was regressed on pretest scores, EC participation, gender, site, grade, race, classroom behavior, IEP, LEP and interaction terms between EC participation and other covariates.

These findings suggest that EC was equally effective across gender, ethnicity, grade, classroom behavior, and English proficiency. However, there were two moderating conditions to note. New York EC students made more gain on the WCJ-Word Attack measure than EC students in Boston and Port Arthur. This differential program impact can be explained by the difference in curriculum. The New York program utilized a curriculum, Book Buddies, which focused more on phonetics. Further, special education students in EC did not benefit from the program as much as non-special ed students in regards to reading comprehension. It is important to note that numerous statistical tests were completed in exploring moderating effects and that the few interactions that were statistically significant could be spurious. Thus these moderating effects need to be interpreted with caution and replication of these findings is warranted.

Exploration of effects of tutoring sessions

Table 6 shows the distribution of tutoring sessions for EC students. The average number of tutoring sessions was 45, with the minimum being 1 session and the maximum 96 sessions. About half of the EC students received between 30 to 49 sessions, and the number of tutoring sessions among students was quite normally distributed. There was no recommended number of sessions that we could use across all programs, given the differences in curricula and the lack of specificity in regards to recommended dosage. Thus, we had no pre-established guideline to determine if students received the full intervention or not. Based on the empirical distribution on the number of sessions received, we chose 35 sessions as the cut-off to indicate if the students received a minimum dose of the intervention. Further, 35 sessions represents about one session a week throughout the program period. Over 75% of the students received at least 35 sessions, and the percentage of students in each category below 35 is between 3% and 8%.

Table 6. Number of tutoring sessions for EC students

Tutoring session	Frequency	Percent
1-9	13	3.6%
10-19	19	5.2%
20-29	30	8.2%
30-34	25	6.8%
35-39	44	12.1%
40-44	44	12.1%
45-49	49	13.4%
50-59	57	15.6%
60-69	52	14.2%
70 or above	32	8.8%
Mean=45.12; Standard Deviation=17.58; Median=45; Range= 1-96		

The findings reported in Table 4 on the statistical differences between the gains made by the EC group and the control students included all of the EC students, even those who received very few tutoring sessions. Yet it is informative to explore program effects with EC student who received more adequate dosages of the intervention. We used a subset of the sample to explore the students who received the full intervention (described above as 35 or more tutoring sessions). Table 7 presents the adjusted posttest means on the reading measures and the accompanying effect sizes. The results show that EC students who had at least a minimum number of sessions made greater

gain than control group students on three of the four reading measures. Effect sizes are .13, .17, and .17, which are larger than those on the full sample presented in Table 4. (A cautionary note is warranted, as students who received more sessions may not be comparable to the control students. That is, a selection bias may have occurred in terms of the subsample of students who received more sessions.)

Table 7. Outcomes for students with 35 or more sessions

Outcome Variable	Treatment Adjusted posttest mean (N=332)	Control Adjusted posttest mean (N=451)	**Program Impact**	**Effect Size**[a]
WJ word attack	96.76 [0.68]	94.59 [0.69]	**2.16* [0.93] t=2.32, p=0.02**	**0.13**
WJ passage comprehension	89.46 [0.54]	87.41 [0.58]	**2.05* [0.80] t=2.56, p=0.01**	**0.17**
PPVT	83.15 [0.45]	82.75 [0.35]	**0.40 [0.56] t=0.71, p=0.48**	**0.03**
Grade-specific reading skills	2.78 [0.03]	2.67 [0.03]	**0.11** [0.04] t=3.10, p=0.003**	**0.17**

Numbers in brackets are standard errors.
a. Standard deviations for the full sample were used to calculate the effect size.

Teacher perceptions of the EC program

Table 8 presents the perception of the EC program provided by 127 teachers who participated in the study. Over 97% of the teachers agreed that EC was beneficial to the students, and the majority rated the program as no or low burden on teachers.

Table 8. Teachers' perceptions of the EC Program (n=127)

To what extent do you agree with the following statement? The EC program is beneficial to the students that participate.	
Extent of agreement	Percent
Strongly agree	59.8%
Agree	37.8%
Neither agree nor disagree	2.4%
Disagree, and strongly disagree	0%

How would you rate the level of burden to teachers of the EC program?	
Level of burden	Percent
No burden	43.3%
Low burden	41.7%
Moderate burden	13.4%
High burden	1.6%

Tutors' perception of the program

Table 9 presents the results from the tutor survey (174 tutors provided ratings on 356 students). This survey included questions about how tutors rated student's progress and the overall quality of their relationship with students during the EC program. The tutors reported that the EC program had a positive impact on students, and their overall relationships with students were good.

Table 9. Tutor perceptions of the EC Program

Question	N	Lowest rating		Middle rating		Highest rating
How much do you feel you helped this student this year? (*not at all* to *a great deal*)	356	6 (2%)	11 (3%)	53 (15%)	123 (34%)	163 (46%)
How much improvement have you seen in this student's reading ability since you began tutoring? (*not much* to *a great deal*)	356	11 (3%)	15 (4%)	63 (18%)	125 (35%)	142 (40%)
How would you rate this student's self confidence today compared to when you first started working with him/her? (*gotten worse* to *improved a lot*)	345	0 (0%)	5 (1%)	34 (10%)	87 (25%)	219 (64%)
How would you rate this student's school behavior today compared to when you first started working with him/her? (*gotten worse* to *improved a lot*)	342	2 (1%)	6 (2%)	84 (24%)	79 (23%)	171 (50%)
How would you describe the overall quality of your relationship with this student? (*poor* to *excellent*)	348	2 (0%)	16 (5%)	45 (13%)	118 (34%)	167 (48%)

Exploring the effect on reading outcomes of quality of the tutoring relationship

We confined our sample to students participating in the EC program to explore whether the quality of the tutoring relationship, as reported by the tutors, affected reading outcomes as assessed by standardized measures and teacher assessment. Results are presented in Appendix G and showed that the relationships between tutors and students were significantly associated with gains made by EC students on two of the four reading measures. A cautionary note is warranted in that students who form good relationships with the tutors may also be inclined toward more positive relationships with all adults, including teachers and parents, and therefore in stronger positions to improve their reading abilities.

Summary and Interpretation of Findings

Students referred to the EC program were very poor readers and clearly in need of assistance. From the scores on the WJ-Passage Comprehension, we can conclude that half of the students referred to EC perform as low as or lower than 84% of the students their age nationwide, and 12% score worse than 97% of the population.

Despite this high level of need, not all the referred students received supplemental assistance. EC had the capacity to serve about half of the referred students, and many control students joined other reading programs (before and after school programs, reading specialist, etc). However, about 30% of the total pool of low-reading students referred to EC did not receive any supplemental reading services over the course of the year. In sum, many students identified as poor readers did not receive any reading assistance outside of normal classroom instruction. EC appears to be a critical part of the network of services available to students who are poor readers.

The students in the EC program made statistically greater gains over the academic year on reading comprehension and on assessments of reading skills made by the teachers ($p<.05$). Additionally, the gains on word attack were marginally significant ($p<.07$). The effect sizes associated with these gains are .10, .13, and .16.

To understand the impact of the EC program, we can compare these effect sizes to those of other and various types of reading interventions. Reading Recovery® (RR) is a one-to-one intensive tutoring program, employing certified teachers specifically trained in the intervention. The What Works Clearinghouse (Institute of Education Sciences, 2007) reports effect sizes around .80. The Tennessee Star program reduced class size to improve academic achievement, and the effect size associated with change in reading scores was .26 (Nye, Hedges, & Konstantopoulos, 2000; Mosteller, 1995). Reading First, a national initiative that promotes instructional practices, did not produce a statistically significant impact on reading comprehension for students in 1st through 3rd grades (Gamse, Jacob, Horst, Boulay, & Unlu, 2008). In this context, the magnitudes of the reading improvements associated with the EC program are substantial, given that the intervention is delivered by trained volunteers.

The EC program succeeded in delivering the intervention to a large number of the students. About half of the EC students received between 30 to 49 sessions, and 76% received over 35 sessions. Although program effects were detected in the full sample, including students who received very few EC sessions, program effects were stronger for the subset of EC students who received 35 or more session (.13, .17, .17). These findings suggest that the EC program would be strengthened by attempts to ensure that all students participate in the program at the intended level.

In general, we did not find evidence to suggest that the program was differentially effective with various subgroups of students. This implies that it is not necessary to target on gender, ethnicity, grade, limited English proficiency, or classroom behavior to maximize program impact. However, findings do suggest that EC students with IEPs, indicating special education, made less improvement than non-special needs students in EC on reading comprehension. The program may benefit from reviewing its approaches to special education students and specifying the curriculum, implementing tutoring training, coordinating with school personnel, and implementing monitoring of student performance.

The finding that the New York EC program had a greater effect on word attack skills is not surprising, given the tutoring curriculum that this site utilizes. However, this finding is useful to remind EC program directors that the tutoring curriculum matters. A review of all curricula used across the EC programs nationwide and their alignment with both program and school district goals may be useful.

Teachers overwhelmingly rated the EC program as beneficial to students, while at the same time, they found that it had no or low burden to them. Although these results derived from teachers' overall perception, they are important findings. If teachers do not have positive perceptions of the program and do not feel that it is worth their effort, program effectiveness and sustainability are threatened.

Tutors perceived that the EC program had a positive impact on students, and their overall relationships with students were good. Further, tutor relationship was related to reading outcomes, with better relationships associated with better outcomes. Clearly, an on-going focus on training tutors to interact in positive ways with the students is important. It is instructive to note that in 18% of the tutor-student matches, the tutors rated the relationship with the student as less than good/excellent. Although a minority of the cases, special support and monitoring of these matches may be warranted.

In sum, these findings indicate that the EC program has statistically significant and substantively important effects on reading outcomes. Further, teachers consider the program to be beneficial to students and a low burden to them.

References

Allison, P. (2006, August). *Multiple imputation of categorical variables under the multivariate normal model.* Symposium conducted at the annual meeting of the American Sociological Association, Montreal, Quebec.

Ballinger, G. A. (2004). Using generalized estimating equations for longitudinal data analysis. *Organizational Research Methods*, *7*(2), 127-150.

Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin*, *107*(2), 238-246.

Browne, N. W., & Cudeck, R. (1993). Alternative ways of assessing model fit. In K. A. Bollen & J. S. Long (Eds.), *Testing structural equation models* (pp. 136-162). Newbury Park, CA: Sage.

Chapman, J. W., & Tunmer, W. E. (2003). Reading difficulties, reading-related self-perceptions, and strategies for overcoming negative self-beliefs. *Reading and Writing Quarterly*, *19*, 5-24.

Eyberg, S., & Pincus, D. (1999). Eyberg Child Behavior Inventory & Sutter-Eyberg Student Behavior Inventory - Revised. Odessa, FL: Psychological Assessment Resources. Retrieved October 1, 2008 from http://www.parinc.com

Gamse, B.C., Jacob, R.T., Horst, M., Boulay, B., and Unlu, F. (2008). *Reading First impact study final report executive summary* (NCEE 2009-4039). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

Gunn, B., Biglan, A., Smolkowski, K., & Ary, D. (2000). The efficacy of supplemental instruction in decoding skills for Hispanic and non-Hispanic students in early elementary school. *Journal of Special Education*, *34*(2), 90-103.

Horton, N. J., & Lipsitz, S. R. (1999). Review of software to fit generalized estimating equation regression models. *American Statistician*, *53*, 160-169.

Institute of Education Sciences. (2007). *What Works Clearinghouse intervention report: Reading Recovery.* Washington, DC: US Department of Education.

Lloyd, M. D., & Dunn, L. M. (1997). Peabody Picture Vocabulary Test Test (3rd ed.). Retrieved October 1, 2008 from http://www.agsnet.com/Group.asp?nGroupInfoID=a12010

Meyer, B. (1995). Natural and quasi-experiments in economics. *Journal of Business & Economic Statistics*, *13*(2), 151-161.

Moerbeek, M., Breukelen, G., & Berger, M. (2003). A comparison between traditional methods and multilevel regression for the analysis of multicenter intervention studies. *Journal of Clinical Epidemiology*, *56*, 341-350.

Mosteller, F. (1995). The Tennessee study of class size in the early school grades. *The future of children: Critical issues for children and youth*, *5*, 113-117.

Nye, B., Hedges, L., & Konstantopoulos, S. (2000). The effects of small classes on academic achievement: The results of the Tennessee class size experiment. *American Educational Research Journal*, *37*, 123-151.

Pajares, F. (2002). Gender and perceived self-efficacy in self-regulated learning. *Theory into Practice*, *41*, 116-126.

Peters, T. J., Richards, S. H., Bankhead, C. R., Ades, A. E., & Sterne, J. A. (2003). Comparisons of methods for analyzing cluster randomized trials: An example involving a factorial design. *International Journal of Epidemiology*, *32*, 840-846.

Rebok, G., Carlson, M., Glass, T., McGill, S., Hill, J., Wasik, B., Ianlongo, N., Frick, K., Fried, L., & Rasmussen, M. (2004). Short term impact of Experience Corps participation on children and schools: Results from a pilot randomized trial. *Journal of Urban Health: Bulletin of New York Academy of Medicine*, *81*, 79-93.

Rubin, D. B. (1987). *Multiple imputation for nonresponse in survey*. New York: John Wiley & Sons, Inc.

Schafer, J. L. (1997). *Analysis of incomplete multivariate data*. New York: Chapman and Hall.

Schafer, J. L., & Olsen, M. K. (1998). Multiple imputation for multivariate missing-data problems: A data analyst's perspective. *Multivariate Behavioral Research, 33*, 545-571.

Schumacker, R. E., & Lomax, R. G. (1996). *A beginner's guide to structural equation modeling*. Hillsdale, NJ: Erlbaum.

Wasik, B.A. (1998). Volunteer tutoring programs in reading: A review. *Reading Research Quarterly, 33*(3), 266-292.

Wayman, J. C. (2003). *Multiple imputation for missing data: What is it and how can I use it?* Symposium conducted at the 2003 annual meeting of the American Educational Research Association, Chicago, IL.

Woodcock, R. W., McGrew, K. S., & Mather, M. (2001). Test review. Woodcock Johnson III Test.Itasca, IL: Riverside Publishing. Retrieved from October 1, 2008 from http://www.riverpub.com/products/wjIIIAchievement/scoring.html

WWC. (2007). *Technical details of WWC-conducted computations* (What Works Clearinghouse Technical Working Papers). San Diego, CA: What Works Clearinghouse.

Yucel, R. M., & Zaslavsky, A. M. (2005). Imputation of binary treatment variables with measurement error in administrative data. *Journal of American Statistical Association*, *100*(472), 1123-1132.

Appendix A: Overview of EC program in Boston, New York and Port Arthur

	New York	Boston	Port Arthur
Student selection	Teacher referral=>ECLAS score	Teacher referral	Teacher referral
Curriculum	Book Buddies	Reading Coaches	Brigance testing and associated work sheets and materials
Dosage	4 times a week, 45 minute sessions for about 24 weeks; does not factor in student attrition and excessive absences	Two times per week for 40 minutes on average; typically served 45+sessions; does not factor in absences.	3 days a week, 25-45 minute sessions
Other reading programs in school	Varies school by school/principal choice: Reading Recovery, Voyager; Teachers College (Columbia)	Varies by school/principal: Reading First's/Harcourt curriculum calls for different interventions including "early reading intervention" etc.; Boston Partners PowerLunch; School Specialists/Reading Recovery;	
Volunteer selection	Interview; application and paper work (including writing sample); reference forms; meet with 2 EC staff; background check	Interview, application, two references, background check- including CORI/SORI	Interview; application and paper work; reference forms; background check
Volunteer training	All volunteers get 32 hours of training which includes intro to program, Book Buddies, lesson plans; new volunteers get an additional 16 hours of training	15-20 hours of training for new volunteers: session/classroom observation(s); 1 hour monthly team meetings, on site practice-specific ½ hour	All volunteers receive 30 hours of pre-service training; new volunteers get an additional 5 hours, team leaders 5 – 10 extra hours.
Volunteer stipend	All volunteers receive stipend: $277 a month for AmeriCorps; $256 per month for no cost volunteers	Stipend levels: non-stipended; Part-time stipend is $185; Full time stipend is $278	All volunteers receive $245 a month. Team leaders receive an extra $60 monthly.
Volunteer hours	16 hours per week	Non-stipended- two or more hours Part-time stipend-10 hours Full-time stipend-15 hours	15 hrs per week

What is the EC staff involvement at the school level?	Participate in school events, parent/ teacher night; sit in on teacher meetings; meet with principals twice a year	Participate in school- and site-sponsored family outreach events; host EC family/community outreach events; meet with principals formally twice a year, phone check-ins	Participate in school- and district-wide activities. Meet with teachers /principals/counselors; other activities as they are presented
What is the EC staff involvement with the tutors?	Training; provide technical assistance; staff person at each site every day with 10-12 tutors, observe daily and evaluate two times per year	Recruit, train, and manage tutors; site visits at least 3x month; on-site coordinator during program operations	Recruit, train (pre- and in-service), monitor at least two times per month
How do teachers become involved?	Talk with principal; principal will have recommendations for teachers who need tutors; seasoned teachers only	Principal designates which grades participate in which; coordinator provides teachers with appropriate referral forms; new teachers will receive information packet	EC staff makes a presentation at the beginning of the school year to the teachers in service trainings.
What records are kept? How is EC participation tracked?	Attendance, lesson plans; log of students progress; tutors turn in time sheets and sign in daily	Attendance sheet (monthly); each session's content is recorded in session plans that are kept in student files; a log is kept of books completed;	Volunteer hours are documented, daily lesson logs are kept per student.
EC established?	1996	1998	1995
How many EC participants?	140-160	305	50
Grades served?	K-2 mainly, up through 5th (classroom assistance)	K-5	K-3
Other/Lead Agencies associated with EC?	Community Service Society	Generations Incorporated	Southeast Texas Regional Planning Commission
Language	Spanish, French, Haitian Creole	Spanish, Cape Verde Creole, Haitian Creole, Vietnamese	English, Spanish, Vietnamese

Appendix B: EC Randomization Summary

	Total Students referred to Treatment Group	Consented Treatment Students		Total Students referred to Control Group	Consented Control Students	
	#	#	%	#	#	%
Boston						
Boston School #1	29	26	90%	52	43	83%
Boston School #2	21	16	76%	28	24	86%
Boston School #3	19	15	79%	29	26	90%
Boston School #4	24	18	75%	40	29	73%
Boston School #5	12	8	67%	9	8	89%
Boston School #6	20	16	80%	13	13	100%
Boston School #7	13	10	77%	16	14	88%
Boston School #8	28	24	86%	16	12	75%
Boston School #9	42	27	64%	31	18	58%
Total:	208	160	77%	234	187	80%
New York						
New York School #1	25	22	88%	33	30	91%
New York School #2	29	22	76%	27	19	70%
New York School #3	40	36	90%	25	22	88%
New York School #4	25	18	72%	39	28	72%
New York School #5	25	21	84%	23	21	91%
New York School #6	25	23	92%	38	31	82%
Total:	169	142	84%	185	151	82%
Port Arthur						
Port Arthur School #1	14	13	93%	13	11	85%
Port Arthur School #2	14	10	71%	9	8	89%
Port Arthur School #3	25	22	88%	12	11	92%
Port Arthur School #4	10	7	70%	10	9	90%
Port Arthur School #5	35	30	86%	40	34	85%
Port Arthur School #6	26	18	69%	27	16	59%
Port Arthur School #7	21	17	81%	20	16	80%
Port Arthur School #8	16	15	94%	12	11	92%
Total:	161	132	82%	143	116	81%
Total:	538	434	81%	562	454	81%

Appendix C: Mathematica Policy Research, Inc. Methodology Report for child and teacher data collection

From September 2006 through June 2008, Mathematica Policy Research, Inc. (MPR) worked closely with researchers from the George Warren Brown School of Social Work at Washington University (WU) to conduct a study of first-, second-, and third-grade children who were eligible to receive tutoring from Experience Corps volunteers. We collected data over a period of two school years. During the first school year, 2006-2007, we collected data at schools in New York City and Boston, and during the second year, 2007-2008, at schools in Port Arthur, Texas. We administered the baseline and follow-up Experience Corps Student Assessment Instrument to all students, with questions customized for their grade level; we asked teachers to complete a baseline and follow-up self-administered Teacher Report on Student Questionnaire for each consented student, a one-time Background and Experience questionnaire about themselves, and a one-time Teacher Review of Experience Corps questionnaire.

In general, we used the same procedures in all sites; the sections that follow will discuss any differences.

A. PRETEST

We conducted 16 pretests of the Experience Corps Student Assessment Instrument during May, June, and July 2006. The Assessment Instrument consisted of a questionnaire on school liking, reading confidence, and the reading and home environment of the child, as well as three reading assessments. The pretests evaluated the placement of the reading assessments within the instrument (either before or after the school liking and reading confidence questions), the use of visual cues for the children, and the wording of questions on reading confidence. We conducted pretests with five first graders, six second graders, and five third graders from a New Jersey convenience sample of low-income children. We administered the complete assessment instrument to 10 of the pretest participants, while six pretests excluded all or some portion of the assessments due to time constraints.

At the beginning of each pretest, an assent script was read to the student. The students then checked the box that corresponded to their decision to participate or not, and wrote their name on the form; all students were able to do so. In one instance a child refused to participate with the survey; all others were agreeable.

Based on the pretest, staff decided to administer the questionnaire at the beginning of the assessment and made wording adjustments to some of the questions.

B. RECRUITING AND TRAINING FIELD STAFF

We recruited, hired, trained, and certified local field staff for administering the student assessment in their districts. All field staff followed the same training protocol; New York City and Boston staff trainings took place in fall 2006 and spring 2007; the Port Arthur trainings took place in fall 2007 and spring 2008.

To conduct the baseline student assessments in the fall, one team leader was hired in each district along with seven field staff from Boston, five field staff in New York City, and eight field staff from the Port Arthur area. In the fall, all field staff attended a two-day training session that focused on the procedures of properly administering several assessments: the Peabody Picture Vocabulary Test

(PPVT IIIA), as well as two tests from the Woodcock-Johnson III (WJ-III) Tests of Achievement, Test 9: Passage Comprehension and Test 13: Word Attack. Following the training, field interviewers conducted practice assessment sessions with children of MPR or Experience Corps staff and friends in order to receive certification to administer student assessments. One field staff member did not pass certification and did not continue on the project. For Boston and New York City staff the fall training took place in Princeton, New Jersey; the Port Arthur fall training took place in Beaumont, Texas.

In the spring, returning field staff participated in a telephone training session in which we reviewed the rules and procedures for conducting the assessments and conducted a read-through of the script. Areas of the assessments that were particularly challenging in the fall were a focus of the training. Staff members received time to practice with one another and team leaders observed all staff members in school when data collection began to ensure that all administration rules were followed. In the spring, returning field staff members included all three team leaders along with three field staff from Boston, four from New York City, and eight from Port Arthur. In order to replace staff that left the project, we recruited, screened, and hired seven new field staff from Boston and New York. New staff attended a two-day training session in Princeton, New Jersey. The two-day training replicated the fall training for new staff. Five of the seven new field interviewers were certified and continued on the project.

C. RANDOM ASSIGNMENT OF STUDENTS AND CONSENT GATHERING

Students in sampled schools were eligible for random assignment and inclusion in the study based on the criteria for Experience Corps tutoring in their districts. In Boston and Port Arthur this was based on teacher recommendation. In New York City, Experience Corps staff administered the Phonological Awareness Literacy Screening (PALS) and used student scores to determine eligibility for tutoring.

We asked schools to recommend twice as many students for tutoring as there were slots available to enable us to randomly assign equal numbers of students to the treatment and control groups.[2] Schools provided MPR with the lists of students who were eligible for tutoring and the number of slots available. We invited all referred students to participate in the study and conducted random assignment on all students nominated for Experience Corps tutoring. We released the results to Experience Corps in conjunction with its scheduled tutoring start dates at each individual school. Participation in the study required explicit parental consent. Consent packets included $2.00 (in Boston and New York City), a letter from WU describing the study, a memo of support from the school principal, a consent form, and a return envelope for the consent form. Materials in the packet were available in six languages; we asked schools which languages to use for their parent population and distributed materials accordingly. We used two methods to distribute consent packets: (1) we mailed or delivered consent packets to the schools and then asked the schools to send the packets home with the students; and (2) for schools that were able to provide address information, we mailed consent packets directly to students' homes by priority mail.

Obtaining high consent rates requires perseverance. In some cases, we made up to five attempts to obtain a completed consent form. Field staff visited the schools to collect returned consent packets

[2] A slot is defined as the time when an Experience Corps volunteer is available to conduct the tutoring program with a student. Volunteers receive training from Experience Corps on how to administer the tutoring program.

and to give the schools new consent packets for students to take home. This often required visiting each classroom and asking teachers to place the forms in students' backpacks. Staff also made regular telephone calls to each school to collect names of newly consented students.

Additional efforts to increase consent return rate included:

1. Some schools permitted MPR to host an in-school party (with healthy snacks, pizza, ice cream, or popcorn and a movie) as an incentive for students to return the consent forms.
2. We enlisted teachers to help collect consent forms and, in return for their assistance, offered each classroom a $25 gift card to purchase books at Barnes & Noble.
3. WU obtained institutional review board (IRB) approval for MPR and/or schools to obtain a verbal consent, as long as two witnesses signed the form and we sent a copy of the signed consent form to the parents.

The overall consent rate for students was 79 percent in Boston, 83 percent in New York City, and 82 percent in Port Arthur.

D. STUDENT DATA COLLECTION

MPR project staff worked with liaisons at each school to arrange the schedule for data collection. We based the timing of the assessment administration visits to schools on the flow of student consents; when we received a batch of consents for a particular school, we assigned a team to go to the school and test the newly consented students. We conducted the student assessments and interviews with students individually in locations assigned to MPR staff by schools. Typical locations for testing were the school library or cafeteria, or Experience Corps office/tutoring space.

After the tester brought a student to the testing location, the first task was to obtain the child's assent. Because of the age of the children and their limited reading skills, the testers read the assent statement to all children, helped the children check the appropriate box, and had them print or sign their names on the form. The text of the children's assent statement is in Figure C.1.

FIGURE C.1
CHILDREN'S ASSENT STATEMENT

> Hi, my name is _____ and I would like to talk to you for a few minutes. I would like to ask you a few questions about school, and see how you are doing with your reading. Your parent(s) know that I will be talking with you. Before we start, I need to get your permission to ask you the questions. Is it OK with you if I ask these questions?

The next task was the administration of the student questionnaire, which measured school liking, reading confidence, and the reading and home environment of the child. For the questions on reading confidence, children looked at cards that had visual depictions of the Likert scales. These progressing pictures illustrated what the response options meant. For example, for the statement, "I am good at reading," the child was shown a card with four boxes. The first box, labeled "not at all true," was empty. The second box, labeled "only a little bit true," contained one star. The third box, labeled "sort of true," had two stars. The fourth box, labeled "very true," had three stars. This series of questions included two training items to allow the children to become familiar with the use of the scales.

Following the student questionnaire, the tester administered the three assessments (PPVT III-A; WJ-III, Test 9; and WJ-III, Test 14). The test session concluded with end-of-scale ratings completed by the tester.

Students were eligible for the baseline assessment if they (1) had parent consent, (2) were still enrolled in the school during the data collection window, and (3) were eligible for tutoring.[3] The same criteria established eligibility for the follow-up assessment, with one clarification: eligibility was also based on remaining enrolled in the school or another school within the district. If a student moved from one school to another in the same district, staff attempted to schedule and complete assessments at the student's new school. A total of 46 students transferred within their school districts over the course of the school year, and field staff completed assessments with 70 percent of the students. An additional 38 students left the school districts and so were ineligible for the follow-up assessment. One student died before the follow-up data collection. Table C.1 describes the response rates for the students.

TABLE C.1

STUDENT ASSESSMENT RESPONSE RATES

District	Baseline	Follow-Up
Boston	100%	99%
New York City	99.5%	96%
Port Arthur	98%	97%

[3] Some students were determined to be ineligible for tutoring based on a standardized test (PALS) after consenting.

E. TEACHER DATA COLLECTION

Prior to the start of data collection at a school, MPR sent questionnaire packets to each teacher with students in the study. Along with study information and a teacher consent form, the packet included up to three types of documents that we asked the teachers to complete if they consented to participate. The first, the Teacher Background and Experience Questionnaire, collected basic demographic information about each teacher. The second, the Teacher Report on Student Questionnaire, was student-specific (one for every consented student) and contained questions on a child's behavior, reading ability, and attendance; teachers completed it at the time of the baseline and follow-up assessments. The third was a short questionnaire about the Experience Corps program (contained only in the spring packet).

The questionnaires were delivered by hand or mailed to teachers at school in a priority mail envelope. In the fall, we sent these questionnaires after receipt of the parent consent; because all consents were not received at the same time, we sent multiple packages (containing the additional consented students' Teacher Report on Student Questionnaires) to the teachers. In the spring we sent the materials in one mailing. Some teachers returned the questionnaires quickly, but others needed multiple reminders to return them. MPR sent WU files containing the names of the teachers who had returned the questionnaires on a weekly basis. WU used these files to process and mail a respondent payment of $15 to teachers for each completed Teacher Report on Student Questionnaire. The total amount a teacher received depended on the number of completed student-level questionnaires. For example a teacher could receive $150 if MPR received 10 completed student-level questionnaires. Table C.2 details the teacher response rates for the Teacher Report on Student Questionnaire.

TABLE C.2

TEACHER REPORT ON STUDENT RESPONSE RATES

District	Baseline	Follow-Up
Boston	77%	80%
New York City	91%	84%
Port Arthur	83%	78%

We received Teacher Background questionnaires from 93 percent of the teachers and we received a completed Teacher Review of Experience Corps Questionnaire from 78 percent of the teachers.

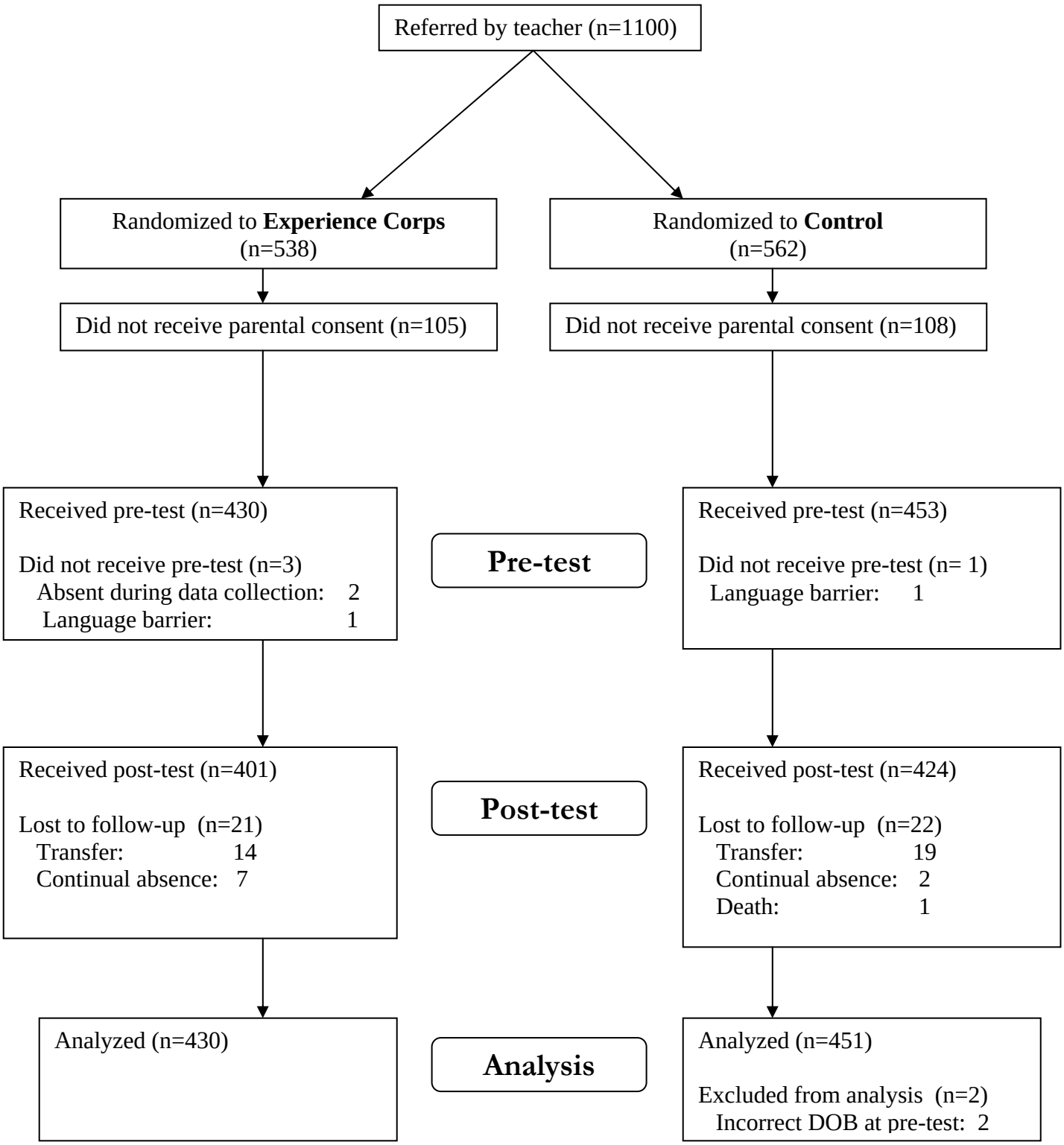
F. SCHOOL RECORDS

In late spring and early summer, following the data collection, MPR requested school records for each consented student, including information on absences, suspensions, race and ethnicity, as well as individual education plan (IEP) status. We collected school records from all participating schools, which represented 90 percent of the consented students (or 99.5 percent of the students who still attended the school at the time of the records collection). Schools were unable to provide full data for the remaining 10 percent of students who transferred out of the school prior to the end of the school year.

G. WEIGHTING

The weights adjust for differences in probabilities of selection both within a school (Boston only) and between schools. The weights are not designed to adjust for attrition or nonresponse. Attrition and nonresponse are defined as students from whom we did not receive consent forms, who left the school after random assignment, or from whom we did not obtain data. The initial weight calculated for each treatment student is the inverse of the probability of selection of being assigned to a tutor. For a control student the initial weight is the inverse of the probability of being a control student (that is, of not being assigned a tutor). The probability of being assigned a tutor is the ratio of the number of tutor slots available and the total number of eligible students. For Boston this was done at the classroom level and for New York City it was done at the school level. For example, if there were five eligible students and two available tutor slots for a classroom in Boston, the probability of being assigned to a tutor is, for each student in the classroom, 2/5 or 0.4; the initial weight for each treatment student in the classroom is then 1/0.4 or 2.5.In the same classroom, the probability of being a control would be 3/5 and the initial weight is 1.6. We adjusted the initial weights so that the sum of the weights was the same for the treatment students and control students in a given school.

Appendix D: Flow Chart of Study Participation



Appendix E: Details of the Statistical Approaches re: Multiple Imputation, Correcting for Clustering, and Calculating Effect Sizes

Multiple imputation for missing observations

There are missing observations across surveys and variables used in the current project. More detailed information on those missing observations is presented in below.

Missing Observations across Surveys and Main Variables

Attrition between pretest and posttest		59
Additional Missing Observations Across Variables		
Teacher Report	Grade-specific reading skills	179
	Teacher's rating of grade level reading performance	187
	Classroom Behavior	179
	Student Motivation	94
	Student Attendance	94
Tutor Survey	Number of Sessions	35
	Tutor Relationship	53
School Record	Race	42
	IEP	78
	LEP	64
	Free Lunch	36
	Days Absent	33

Missing observation were imputed based on a Markov Chain Monte Carlo (MCMC) multiple imputation. Multiple imputation procedure replaces each missing observation with a set of predicted values using existing values from other variables, and these multiply imputed values represent the uncertainty about the right value to impute (Rubin, 1987; Schafer, 1997; Wayman, 2003). Standard statistical analyses are then performed for each imputed dataset, and the results are combined to produce an overall result. Through this procedure, statistically valid inference is possible since it is proven that overall estimates from multiply imputed data will be unbiased, properly reflecting uncertainty due to missing observations (Wayman, 2003).[4]

There are assumptions for multiple imputation. First, MAR (Missing At Random) condition should be satisfied. Second, MCMC multiple imputation, a most widely used approach, needs an assumption of linearity and multivariate normality. However, simulation studies have found that multiple imputation is robust to departures from these assumptions (Wayman, 2003; Yucel & Zaslavsky, 2005).

There are several approaches to conduct multiple imputation. Among others, we mainly used MCMC multiple imputation for the current study, which is one of the most widely used methods for

[4] Contrary to multiple imputation, single imputation does not reflect uncertainty about the right value to impute, and the resulting estimated variances of the parameter estimates will be biased (Yang, n.d.)

arbitrary missing data. Assuming that the data are from multivariate normal distribution, MCMC multiple imputation uses the EM algorithm and the method of generating random draws from probability distribution via Markov chains (Schafer, 1997). Our data include a set of categorical variables, so it is possible that linearity and multivariate normality assumptions are violated. However, as mentioned above, MCMC approach is robust to departures from these assumptions.[5]

One of the issues in using MCMC method for categorical data is whether imputed values should be rounded or not. It has been common practice to impute categorical data using MCMC method and round them. However, recent studies found that the practice of rounding may produce biased estimates (Allison, 2006). Based on these findings, we conducted MCMC multiple imputation without rounding for the current study.

The number of imputation is another important issue in multiple imputation. Guidelines for the number of imputation vary by studies and characteristics of data such as a proportion of missing values. We created five imputed datasets for the current study following suggestions by Schafer & Olsen (1998). Given the fraction of missing data in our study, five imputations may be efficient enough.

The following table shows the variability added due to multiple imputation.

Variable	Relative increase in variance
Posttest word attack	0.040
Posttest passage comprehension	0.068
Posttest PPVT	0.023
Posttest grade-specific reading skill	0.137
Hispanic origin	0.008
Other race	0.105
Classroom behavior	0.113
IEP	0.163
LEP	0.029
Number of Sessions	0.072
Tutor relationship	0.010

[5] Another popular method for multiple imputation is a MICE (Multiple Imputation by Chained Equation) approach. One of the advantages of MICE is it does not need an assumption of multivariate normality. We also analyzed data with MICE method, and have substantially similar results.

The following table corresponds to Table 2 in the text. In the text, the sample is described using raw, unimputed data; therefore the sample size depends on the specific variable. Here, the sample description is presented on the imputed data; thus, sample size is the same across all variables. .

Sample description at pretest (Data with Multiple Imputation)

	Total (N=881)	EC (N=430)	Control (N=451)
Reading Outcomes			
WJ Word Attack (Standardized Score)	91.89 (20.45)	91.89 (20.31)	91.89 (20.61)
WJ Passage Comprehension (Standardized Score)	84.41 (13.72)	83.99 (13.59)	84.84 (13.85)
PPVT (Standardized Score)	80.95 (13.98)	80.87 (14.30)	81.03 (13.65)
Grade-specific reading skills	2.37 (0.59)	2.36 (0.57)	2.38 (0.62)
Demographics			
Gender			
Male	451 (51%)	209 (49%)	242 (54%)
Female	430 (49%)	221 (51%)	209 (46%)
Race			
African American	511 (58%)	259 (60%)	252 (56%)
Hispanic Origin	321 (36%)	145 (34%)	176 (39%)
Others	49 (6%)	26 (6%)	23 (5%)
Grade			
1st grade	363 (41%)	180 (42%)	183 (40%)
2nd grade	318 (36%)	162 (38%)	156 (35%)
3rd grade	200 (23%)	88 (20%)	112 (25%)
Age	7.09 (1.10)	7.07 (1.07)	7.12 (1.14)
School Events			
Free lunch			
Yes	829 (94%)	401 (93%)	428 (95%)
No	52 (6%)	29 (7%)	23 (5%)
IEP (Individualized Education Plan)			
Yes	131 (15%)	64 (15%)	67 (15%)
No	750 (85%)	366 (85%)	384 (85%)
LEP (Limited English Proficiency)			
Yes	207 (23%)	97 (23%)	110 (24%)
No	674 (77%)	333 (77%)	341 (76%)
Student Behaviors			
Classroom Behavior	3.58 (0.76)	3.57 (0.75)	3.59 (0.77)

Adjusting clustering effect

The data used in this project have a nested or hierarchical structure (e.g., students are nested within classrooms, classrooms are nested within schools). In these clustered data, outcomes of individuals within the same cluster are likely to be correlated; therefore, the assumption of OLS, independence of observations, is violated. A failure to incorporate within-cluster correlations into the analytic model leads to incorrect standard errors and p-values (Ballinger, 2004; Peters et al., 2003).

There are several statistical options to deal with this issue, including GEE (Generalized Estimating Equation), Multi-level modeling (or HLM), and cluster robust standard errors (or Huber-White sandwich estimator). We use GEE in this project for the following reasons. First, with a large enough sample size, GEE provides correct parameter estimates and standard errors even though the correlation matrix is misspecified (Ballinger, 2004; Moerbeek et al., 2003). Second, GEE does not require the assumption of multivariate normal distribution (Ballinger, 2004).

There are two possible cluster variables: teacher (classroom) and school, and we use "teacher" as a clustering unit. First, according to the variance component analysis, variance among teachers tend to be larger compared to that among schools for the most outcomes. Secondly, considerable literature suggests that there should be an enough number of clusters for the GEE estimates to be valid. Horton and Lipsitz (1999) suggests that GEE estimates should be used with more than 20 clusters.

In the GEE model, we specify an exchangeable working correlation matrix where within-cluster observations are assumed to be equally correlated. Also, analysis of GEE parameter estimates are based on the empirical standard error estimates (not model-based standard error estimates) because they are robust to misspecification of working correlation matrix.

Effect size

In this study, Hedge's G statistics are used to compute effect sizes. The formula is as follows:

$$\text{Hedge's g} = \frac{X'_1 - X'_2}{\sqrt{\frac{(n_1 - 1)S_1^{\,2} + (n_2 - 1)S_2^{\,2}}{(n_1 + n_2 - 2)}}}$$

where X'_1 and X'_2 are adjusted posttest means, n_1 and n_2 the sample sizes, and S_1 and S_2 the student-level unadjusted posttest standard deviations for the EC group and the control group, respectively (WWC, 2007).

Appendix F: Effects of EC Program, Controlling for Covariates

	WJ Word Attack	WJ Passage Comprehension	PPVT	Grade-specific reading skills
Scores in pretest	0.42***	0.42***	0.68***	0.71***
	[0.04]	[0.04]	[0.03]	[0.04]
EC	**1.59†**	**1.52***	**0.03**	**0.10****
	[0.88]	**[0.73]**	**[0.51]**	**[0.03]**
Female	-2.25†	-0.26	-1.20*	0.06†
	[1.23]	[0.78]	[0.54]	[0.04]
Race				
(African-American)				
Hispanic Origin	0.60	-0.56	-1.65*	0.03
	[1.70]	[1.10]	[0.68]	[0.06]
Other Race	-2.79	-0.12	3.66**	0.12
	[2.50]	[1.56]	[1.11]	[0.11]
Grade				
(2st grade)				
1st grade	4.99***	3.52***	0.93	0.16**
	[1.10]	[0.90]	[0.65]	[0.05]
3rd grade	1.80	-1.08	-1.45†	-0.03
	[1.26]	[0.84]	[0.80]	[0.05]
Site				
(New York)				
Boston	-1.51	0.18	2.38***	-0.0004
	[1.42]	[0.96]	[0.72]	[0.05]
PA	-0.10	1.45	0.30	-0.13*
	[1.11]	[0.96]	[0.73]	[0.05]
Classroom Behavior	3.02***	2.06**	1.09*	0.04
	[0.83]	[0.69]	[0.44]	[0.02]
IEP	-4.34*	-4.71***	-2.63**	-0.12*
	[1.69]	[1.19]	[0.86]	[0.06]
LEP	-0.73	-1.90†	-0.81	-0.04
	[1.49]	[1.08]	[0.74]	[0.07]

Note: N=881; † p<.10, * p<.05, ** p<.01, *** p<.001

Appendix G: Effect of Quality of the Tutoring Relationship (N=430; Only EC students)

	WJ Word Attack	WJ Passage Comprehension	PPVT	Grade-specific reading skills
Scores in pretest	0.35***	0.43***	0.69***	0.66***
	[0.06]	[0.05]	[0.03]	[0.06]
Tutoring Relationship	**0.69**	**2.02****	**0.99***	**0.02**
	[0.81]	**[0.67]**	**[0.48]**	**[0.03]**
Female	-2.62	-1.03	-1.41†	0.04
	[1.56]	[0.86]	[0.83]	[0.05]
Race				
(African-American)				
Hispanic Origin	2.17	-0.15	-2.98**	0.03
	[1.97]	[1.25]	[0.92]	[0.10]
Other Race	-4.69	-1.04	3.14*	0.11
	[3.45]	[2.02]	[1.58]	[0.15]
Grade				
(2st grade)				
1st grade	6.14***	4.16***	0.74	0.19**
	[1.91]	[1.01]	[0.93]	[0.06]
3rd grade	0.07	-1.49	-1.60	0.04
	[1.91]	[1.23]	[1.16]	[0.07]
Site				
(New York)				
Boston	-4.05*	-0.70	1.02	-0.04
	[1.67]	[1.17]	[1.07]	[0.08]
PA	-1.60	-0.21	-1.09	-0.18*
	[1.58]	[1.24]	[0.04]	[0.08]
Classroom Behavior	4.15***	2.32**	1.00†	0.04
	[0.96]	[0.73]	[0.62]	[0.04]
IEP	-5.53**	-6.33***	-2.37†	-0.12
	[2.10]	[1.62]	[0.96]	[0.09]
LEP	-2.72	-2.65†	-0.41	-0.04
	[2.17]	[1.46]	[1.05]	[0.10]

Note: † p<.10, * p<.05, ** p<.01, *** p<.001