



Education Innovation and Research EVALUATION DESIGN PLAN¹

Grantee Name/ID: Tufts University [U411C190006]

¹ Price, C., Goodson, B., Wolf, A. (2018). *Evaluation Design Plan Template*. Prepared for the Institute of Education Sciences, U.S. Department of Education, Washington, D.C. as part of contracts to provide evaluation technical assistance for evaluations of interventions funded by the Investing in Innovation, First in the World, and Education Innovation and Research programs.

Table of Contents

1. Evaluator information.....	3
1.1. Contact Information.....	3
1.2. Confidentiality protection.....	3
1.3. Independence of evaluation	3
2. Summary of Intervention(s).....	5
3. Impact Evaluation.....	8
3.1. Impact Study	8
3.1.1. Introduction to Impact Study	8
3.1.2. Research questions	9
3.1.3. Control or comparison condition.....	9
3.1.4. Sample identification, selection and assignment.....	10
3.1.5. Key measures and plan for collecting data	13
3.1.6. Attrition (RCTs only)	16
3.1.7. Statistical analysis of impacts.....	19
3.1.8. Baseline equivalence testing – QEDs, RCTs with high attrition, cluster RCTs with potential bias from entering individuals.....	21
4. Implementation Evaluation	23
4.1. Implementation Study of the EIR-funded Intervention.....	23
4.1.1. Logic model for the intervention	23
4.1.2. Research questions for evaluation of implementation	25
4.1.3. Measuring fidelity of implementation	26
4.1.4. Data collection plan.....	28
4.1.5. Fidelity reporting plan	28
4.1.6. Plan for providing performance feedback to inform project design (Early-phase grants only)...	29
5. Alignment of Evaluation Samples with Units Receiving the Intervention.....	31
6. References	33

1. Evaluator information

1.1. Contact Information

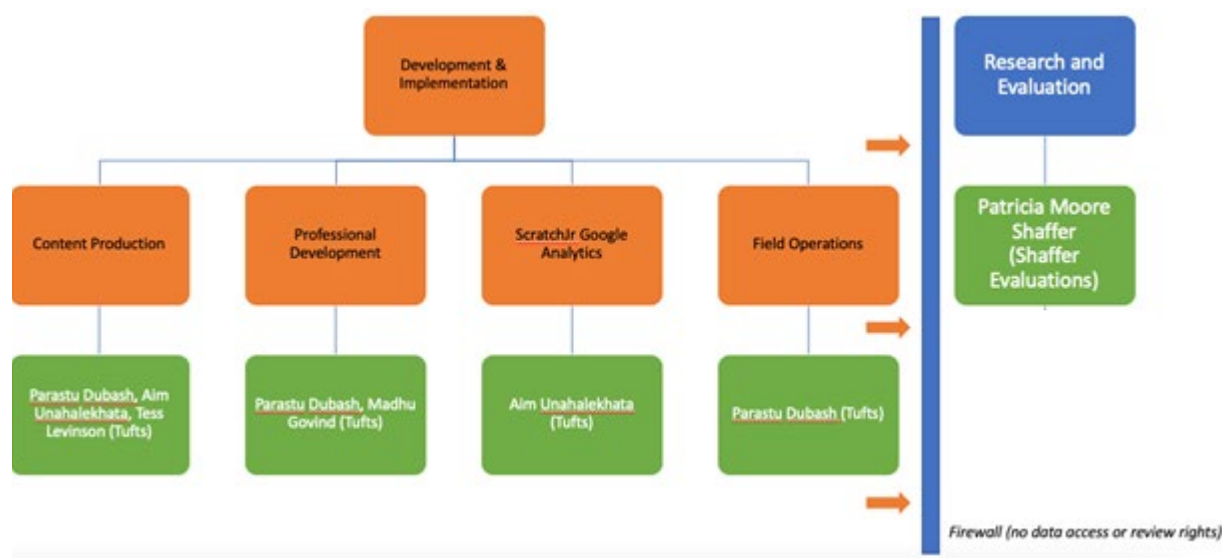
- 1) Patricia Moore Shaffer, Shaffer Evaluation Group, 1769 Jamestown Road, Ste 117, Williamsburg, VA 23185; patricia.shaffer@shafferevaluation.com; 703-582-9749
- 2) Marina Bers, DevTech Research Group, Tufts University, Eliot-Pearson Department of Child Study and Human Development, 105 College Ave., Medford, MA, 02155; marina.bers@tufts.edu; 617-627-3396

1.2. Confidentiality protection

Tufts has internal IRB processes that uphold rigorous standards with respect to the protection of human subjects. IRB approval has been obtained. All sensitive data will be securely stored and handled. Identifiable personal data will be accessible only to staff working directly on the project. No individuals will be identified in any reports.

1.3. Independence of evaluation

The evaluation is being led and conducted by Shaffer Evaluation Group. Shaffer Evaluation Group is working together with the Research and Evaluation team of the DevTech Research Group, which is working independently from the intervention development and implementation team. A firewall is implemented between the Tufts Research and Evaluation team of the DevTech Research Group and the intervention development and implementation team (see illustration). All decisions regarding assignment, data collection, data analysis, and final reporting are made by Shaffer Evaluation Group.



Assignment

Random assignment of the schools to either the treatment or the comparison group will be done by Shaffer Evaluation Group in consultation with the Research and Evaluation team of the DevTech Research Group.

Data collection

Student outcome data are obtained in two ways:

1. Scores from district or state achievement tests (such as i-Ready, MAP Fluency, DIBELS) are provided directly to Shaffer Evaluation Group by the schools in Rhode Island and Boston Public Schools using an upload process to a secure, cloud-based platform (Box.com).
2. Researcher-developed instruments (TechCheck, Coding Stages Assessment) are virtually administered by research assistants from Shaffer Evaluation Group. All data collection will be conducted online using a secure Zoom account. Assessment data are entered by research assistants into a Qualtrics online form; the Qualtrics forms are owned and managed by Shaffer Evaluation Group. Shaffer Evaluation Group staff download assessment data from the secure Qualtrics server.

Data storage and access

The raw data are stored and time-stamped on a secure server to which only staff of Shaffer Evaluation Group has access. A copy of the data is stored on a secure server to which both Shaffer Evaluation Group and the Research and Evaluation team of the DevTech Research Group have access to enable collaboration on the analyses. No one from the intervention development and implementation team has access to this server. If anyone from the intervention

development and implementation team wants to conduct research using the data, they will be given a copy (in line with IRB requirements). This research will be unrelated to the analyses that are done for evaluation purposes (see data analysis below).

Data analysis

Shaffer Evaluation Group and the Research and Evaluation team of the DevTech Research Group work together to select statistical methods that are suitable for measuring the effect of the intervention. The analyses will be conducted by the Research and Evaluation team of the DevTech Research Group and audited by Shaffer Evaluation Group. All analyses will be done using open and reproducible tools such as scripts and R Notebooks, which allow retracing and checking every step (such as removal of duplicate or incorrect data, recoding of variables, or model specifications). This way the analyses can be re-run using the original raw data. No one from the intervention development and implementation team, including the PI, will be involved in the data analysis. Only Shaffer Evaluation Group and the Research and Evaluation team of the DevTech Research Group will have access to the analysis scripts until the final report has been submitted.

Reporting

The estimated impact will be reported by Shaffer Evaluation Group in consultation with the Research and Evaluation team of the DevTech Research Group.

2. Summary of Intervention(s)

The objective of the intervention is to improve Computational Thinking (CT) skills, coding skills, and language arts comprehension outcomes in early elementary school students, including in schools serving disadvantaged students, using the technology platform ScratchJr. An important short-term outcome of the intervention (mediator) is to improve pedagogical and content knowledge in teachers.

To achieve this, a new curriculum, CAL (Coding as Another Language) and affiliated teacher professional development will be implemented in kindergarten through 2nd grade in a sample of elementary schools in a high-poverty, urban school district in the United States.

The curriculum builds on DevTech's previously developed pilot units (Bers, 2018) and will be aligned with the K-12 CS Framework (K-12 CS Framework Steering Committee, 2016) and the Standards for Technological Literacy (International Technology and Engineering Education Association, 2007), as well as Common Core Frameworks for Math and Literacy (National Governors Association Center for Best Practices & Council of Chief State School Officers, 2010), and Virginia Department of Education's Standards of Learning for English and Standards of Learning for CS (Virginia Department of Education, 2017).

The CAL curriculum is organized around powerful ideas that are fundamental to computational thinking and, at the same time, are developmentally appropriate for young children. The curriculum introduces coding and computational thinking in a playful, developmentally appropriate way by integrating powerful ideas of computer science with literacy skills.

Students will receive a total of 24 lessons of 45 minutes each.

The Coding as Another Language intervention will be implemented in a delayed treatment design. Following a pilot study in Year 2, in Year 3, Group 1 (treatment group) will be trained in the delivery of the Coding as Another Language curriculum and provided resources and support to implement the curriculum during a 12-week period (Fall 2021). Group 2 schools (control group) will deliver business-as-usual during Year 3 and delay implementation of the intervention until Year 4; Group 2 teachers will not have access to training or the curriculum. Implementation of the intervention in Group 2 in Year 4 will be used to pilot a scale-up study. Note that Year 4 is not part of the evaluation.

For the Year 3 study (Impact Study), the DevTech Research Group will train the teachers delivering the curriculum in Group 1 schools. The impact of the intervention will be measured by comparing student outcomes in the treatment and the control group. Student outcomes will be measured before and after the administration of the curriculum/business-as-usual.

Although the scale-up study in Year 4 is not part of the evaluation, it has consequences for the randomization. To test the scalability of the intervention, DevTech will train a group of Tech Leaders, who will become coaches to teachers in Group 2 schools, starting in 2021. The training duration is one year (fall term 2021, spring term 2022, summer term 2022). These Tech Leaders are classroom teachers, librarians and media/technology specialists at schools in Rhode Island and Boston Public Schools (BPS). Please see section 3.1.4.2. (Identification/selection of schools) for details.

The study target population is K-2 students and teachers in Rhode Island and in Boston (Massachusetts). Both sites struggle with poverty. Rhode Island ranks among the lowest performing states in the US in terms of education, 75% of all students come from economically disadvantaged backgrounds. The student population at BPS is 14% Caucasian, 33% Black, 42.5% Hispanic, 9% Asian, and 1.5% other/multi-race; currently 72% of BPS students are economically disadvantaged.

	Outcome Domain	Name of Outcome Measure (and subtest)
Outcome 1	Computational thinking	TechCheck
Outcome 2	Coding knowledge	Coding Stages Assessment

Outcome 3	Comprehension	<ul style="list-style-type: none">• DIBELS – subtest: MAZE• MAP Fluency – subtests listening comprehension and picture vocabulary• STAR Reading – literary text• iReady – comprehension: literature• AimsWeb Plus – reading comprehension• Independent Reading Level Assessment (IRLA) – reading comprehension• Fastbridge Reading Screener
-----------	---------------	---

3. Impact Evaluation

Impact studies described in Chapter 3	
Title	Notes
Impact Study A: Rhode Island School RCT	Impact of the CAL curriculum intervention on 2 nd grade student outcomes (schools randomly assigned to conditions)
Impact Study B: Boston Public Schools RCT	Impact of the CAL curriculum intervention on 2 nd grade student outcomes (schools randomly assigned to conditions)

3.1. *Impact Study*

3.1.1. Introduction to Impact Study

The two impact studies will use a randomized control trial design that assigns schools to either the treatment or to a control condition. Teachers in treatment schools will be trained to deliver the intervention, so that the intervention is delivered at the classroom level. Groups 1 (treatment) and 2 (control) will be tracked during Year 3. Group 1 will implement the Coding as Another Language (CAL) curriculum supported by training from the DevTech Research Group during the first year of implementation. Group 1 and 2 students will be compared using three-level hierarchical linear modeling (HLM), controlling for covariates at the student, teacher, and school levels, to test for differences in the following outcomes: computational thinking, coding skills, and early language skills. Group 2 schools will implement the CAL curriculum in Year 4.

The research questions, outcomes, baseline measures, sample, and contrasts relevant to the study being described in this section are listed in “2000917_Early52_contrast_tool.xlsm”

3.1.2. Research questions

Research Question 1	What is the impact of three months of CAL curriculum on second grade student development of computational thinking compared to the business-as-usual condition?
Research Question 2	What is the impact of three months of CAL curriculum on second grade student development of coding skills compared to the business-as-usual condition?
Research Question 3	What is the impact of three months of CAL curriculum on second grade student development of reading comprehension compared to the business-as-usual condition?

3.1.3. Control or comparison condition

3.1.3.1. Identification/selection of study districts

The comparison condition is an active control, where students receive “business-as-usual” instruction. Business-as-usual varies from district to district.

Study A: At Rhode Island, 2nd graders typically do between 1.5 and 3 hours of classroom work each day, using both block and integrated models. Literacy activities include read-aloud, guided reading, and shared reading with fluency, phonemic awareness and comprehension work, workstations, shared writing, reading and writing workshops.

Study B: At BPS, 2nd grade students typically do around 3 hours of classroom work each day, with the following components:

- **Text Talk** (45 minutes/day) Whole group conversations and work centered on a text—read aloud or other format (video, photograph, quote, article)—and embedded vocabulary lessons.
- **Foundations** (30 minutes/day) Systematic teaching of foundational skills in reading and spelling, emphasizing phonemic awareness, phonics, word study, high-frequency words, fluency, vocabulary and handwriting.
- **Stations** (45 minutes/day) Responsive literacy instruction, including independent and collaborative practice of explicit literacy skills: Guided Independent Reading and Listening and Speaking, Vocabulary, and Word Work Stations. During this time, teachers convene small groups for guided reading and other strategic instruction.
- **Science and Engineering** (30 minutes/day, twice a week) Science and engineering lessons related to the content of the unit and addressing Science and Engineering Standards.

- **Studios** (30 minutes/day, three times a week) Exploration of weekly questions and texts in a variety of media. Studios include Art, Building, Discovery, Research, and Writing and Storytelling.
- **Writing** (30 minutes/day) A writing block grounded in Systemic Functional Linguistics (SFL). Children write in various genres for specific audiences, in the context of the unit content.
- **Storytelling/Story Acting** (embedded and during Studios; about 10 minutes, twice a week) A system of telling, writing, and acting out personal stories.

Note that there are currently no plans to include CS topics in the control group in either study. Some schools/teachers may decide to cover CS topics in their individual teaching plans, but there are no systematic efforts in place.

3.1.4. Sample identification, selection and assignment

3.1.4.1. Identification/selection of study districts

School districts with racial/ethnic and economic diversity were identified and sought for this study.

Study A: Rhode Island Department of Education has previously made efforts to integrate computer science in all grades through its CS4RI initiative. The state's schools include economically disadvantaged schools that will be part of the study. For the study, individual schools from several Rhode Island school districts will be involved.

Study B: Boston Public Schools offer strong racial/ethnic and economic diversity, with 42.4% Hispanic, 30% Black, 14.9% White, 9% Asian, and 3.7% other/multiracial student population and 73.2% of students classified as economically disadvantaged (2020-21).

3.1.4.2. Identification/selection of study schools

Study A: At Rhode Island, schools will be invited through calls by the Rhode Island Department of Education (RIDE) to submit applications to participate. Title I schools will be particularly encouraged to apply. Participation is voluntary. Criteria for selection include demographics (Title I schools will be given preference) and number of participating classrooms (schools with more participating classrooms will be given preference). The schools will be randomized with blocks. Blocks will be formed using three criteria:

- 1) Poverty quartile
- 2) Number of estimated participating students
- 3) Teachers who are part of the Tech Leader group only

The following blocks will be formed using the first two criteria in order to ensure comparable demographics and comparable sample sizes across the treatment and the control group: 1) schools in the lowest and second-to-lowest poverty quartile that expect fewer than 100 students to participate 2) schools in the lowest and second-to-lowest poverty quartile that expect more than 100 students to participate, 3) schools in the highest and second-to-highest poverty quartile.

Further, the larger research project will also contain a pilot for a scale-up study. For this, 11 Tech Leaders, selected from a diverse group of teachers, coaches, and technology coordinators and integrators in Rhode Island, will be trained so that they can continue the implementation of the curriculum. Tech Leaders will enroll in the Early Childhood Technology (ECT) graduate program at Tufts University, which blends online courses with one-week intensive summer supervised practicum at the lab school: the Eliot-Pearson Children's School. They also will participate in curriculum-specific training conducted by the DevTech Research Group. During Year 3, DevTech will train and coach group 1 along with Tech Leaders. Later on, during Year 4, Tech Leaders will train and coach group 2. To avoid potential bleed between the treatment and the control condition (i.e., Tech Leaders' professional development as coaches may affect their teaching or teaching at the school where they work), whether or not a school has teachers who are Tech Leaders will be taken into account when blocking. In Rhode Island, all participating schools will be associated with a Tech Leader (see below), so the forming of blocks is not necessary.

Study B: At BPS, central office will invite schools to submit applications to participate. Title I schools will be particularly encouraged to apply. Participation is voluntary. Criteria for selection include demographics (Title I schools will be given preference) and number of participating classrooms (schools with more participating classrooms will be given preference). Similar to Rhode Island, the schools will be randomized with blocks. Blocks will be formed using three criteria:

- 1) Poverty quartile
- 2) Number of estimated participating students
- 3) Teachers who are part of the Tech Leader group only

The following blocks will be formed using the first two criteria in order to ensure comparable demographics and comparable sample sizes across the treatment and the control group: 1) schools in the lowest and second-to-lowest poverty quartile that expect fewer than 100 students to participate 2) schools in the lowest and second-to-lowest poverty quartile that expect more than 100 students to participate, 3) schools in the highest and second-to-highest poverty quartile.

Further, the larger research project will also contain a pilot for a scale-up study. For this, 9 Tech Leaders, selected from a diverse group of teachers, coaches, and technology coordinators and integrators in BPS, will be trained so that they can continue the implementation of the

curriculum. Tech Leaders will enroll in the Early Childhood Technology (ECT) graduate program at Tufts University, which blends online courses with one-week intensive summer supervised practicum at the lab school: the Eliot-Pearson Children's School. They also will participate in curriculum-specific training conducted by the DevTech Research Group. During Year 3, DevTech will train and coach group 1 along with Tech Leaders. Later on, during Year 4, Tech Leaders will train and coach group 2. To avoid potential bleed between the treatment and the control condition (i.e., Tech Leaders' professional development as coaches may affect their teaching or teaching at the school where they work), whether or not a school has teachers who are Tech Leaders will be taken into account when blocking.

For BPS, the number of schools has not been determined at the time of writing. If the situation is the same as in Rhode Island (i.e., a Tech Leader at each school), blocking will not be necessary. If there are schools with and schools without Tech Leaders, 6 blocks will be formed: 1) schools in the lowest and second-to-lowest poverty quartile that expect fewer than 100 students to participate and that have a Tech Leader, 2) schools in the lowest and second-to-lowest poverty quartile that expect fewer than 100 students to participate and that do not have a Tech Leader, 3) schools in the lowest and second-to-lowest poverty quartile that expect more than 100 students to participate and that have a Tech Leader, 4) schools in the lowest and second-to-lowest poverty quartile that expect more than 100 students to participate and that do not have a Tech Leader, 5) schools in the highest and second-to-highest poverty quartile that have a Tech Leader, 6) schools in the highest and second-to-highest poverty quartile that don't have a Tech Leader. There will thus be either 3 or 6 blocks from which schools will be randomly allocated to the treatment and the control condition.

3.1.4.3. Identification/selection of study teachers

Study A and B: Grade 2 teachers at the schools in both studies will participate on a voluntary basis in the study; teachers volunteered for the study prior to random assignment of schools. It is possible that teachers join the clusters after randomization but participating classrooms will remain intact. For example, if a teacher is moved to another grade or school after randomization, the new teacher assigned to their classroom is considered to be part of the study. We judge the risk of bias associated with these joiners low. We do not expect that teachers will join the treatment clusters later that differ systematically from those who entered the comparison cluster. The intervention is only 3 months in total and unlikely to be a reason for teachers to join a Group 1 school, especially given that in the following year, all remaining schools will implement the curriculum. We have received consent forms from study A and is collecting consent forms from study B. We expect ~51 teachers to be included in the two studies (Study A: ~9 in the treatment group, ~12 in the control group; Study B: ~15 in the treatment group, ~15 in the control group).

Study A only: The CAL curriculum will be implemented using different models. In some RI schools, the 2nd grade classroom teacher may teach the curriculum with the support of English Language Learners (ELL) and Special Education (SPED) co-teachers. In some schools, a STEAM teacher may teach the curriculum with or without the support of the classroom teacher.

In RI, all participating teachers – classroom, ELL, SPED, and STEAM teachers will participate in CAL professional development. The evaluator will gather detailed data on CAL implementation models in each participating school.

3.1.4.3. Identification/selection of study classes

Studies A and B: All grades K-2 classes will be included in the intervention (but not necessarily all classes at a given school), but only Grade 2 will be the focus of the evaluation. Grade 2 was selected for study because standardized assessments are available for literacy in this grade. Classrooms led by teachers who volunteered for the study will be included in control or treatment groups.

3.1.4.4. Identification/selection of students

Study A: Students were assigned to classrooms after random assignment by school administrators following their own school schedules and methods. Currently, more classrooms (12) are available in the control group than in the treatment group (9). To reduce cost, we are planning to randomly subsample students from all available classrooms in the control group to have about the same size in the treatment group. Assuming an effect size of 0.15, $\alpha=0.05$, power of 0.8, number of tested predictors of 2 (time and condition), the needed sample size for linear multiple regression (fixed model) is 68. We estimated the design effect of 1.77 using a method from Rutterford, Copas, and Eldridge (2015) when cluster sizes vary and sample 120% more for any possible attrition, the sample size at the individual student level becomes 145. With 21 classrooms, we will have plenty sample size to ensure the power for our analyses.

It is highly likely that there will be students that enter the schools after randomization. The study sample will be defined as students who were in Grade 2 participating classrooms prior to the start of curriculum implementation. If a student joins the class after the start of the curriculum, they will receive the curriculum, but their data would be excluded from the analysis.

Study B: All students in participating Grade 2 classrooms will be included. Students were assigned to classrooms before random assignment.

It is highly likely that there will be students that enter the schools after randomization. The study sample will be defined as students who were in Grade 2 participating classrooms before randomization. If a student joins the class after the start of the curriculum, they will receive the curriculum, but their data would be excluded from the analysis.

3.1.5. Key measures and plan for collecting data

Both Studies: Three domains of student outcomes will be measured: computational thinking, coding knowledge, and literacy comprehension. The baseline measures correspond with the outcome measures (i.e., the same assessments/subtests will be used). Computational thinking will be assessed using the validated instrument TechCheck, a developmentally appropriate

assessment in an “unplugged”, multiple-choice format (Relkin, de Ruiter & Bers, 2020). TechCheck has a reported internal consistency of Cronbach’s $\alpha = .68$. Coding knowledge will be assessed using the ScratchJr Coding Stages Assessment (CSA) (De Ruiter & Bers, n.d.). This CSA is one-on-one and involves a series of tasks and questions mapped onto each of the six coding stages (Bers, 2019), using the ScratchJr app. The CSA ScratchJr has a split-half reliability of $\lambda_6 = .94$. Significant improvement in the treatment group is expected in both TechCheck and CSA.

TechCheck and the Coding Stages Assessment will be administered by trained research assistants. The external evaluator will select, train, and supervise the research assistants. The training (but not the data collection) will be conducted with support from the Tufts intervention development and implementation team.

Reading comprehension outcomes will be collected by Rhode Island schools and BPS following standardized procedures normally implemented.

In addition to these baseline and outcome measures, the following data will be collected to be included as covariates in the model:

- Treatment indicator: a variable indicating the group to which a school was randomly assigned (0 = control group, 1 = CAL).
- Race: a categorical variable indicating whether the student is African American, Caucasian, Hispanic, American Indian/Alaska Native, Asian, Native Hawaiian/Pacific Islander or Multiracial.
- Gender: a categorical variable indicating if the student is male, female, or doesn’t prefer to provide their gender
- Disability: a binary variable indicating whether or not a student has a recognized disability (0 = no, 1 = yes)
- ELL: a binary variable indicating whether or not a student is registered as being an English Language Learner (0 = no, 1 = yes)
- Lunch: A continuous variable indicating the schools’ percentage of free/reduced lunch

Study A only: In the literacy domain, Rhode Island schools are using different assessments. We will choose those subtests that are the closest match to measuring comprehension or use the primary scale. In the table below, we are listing the assessments (and for some the corresponding subtests) that we are aware the participating schools are using. At the time this plan was submitted, we were still collecting information on available assessments. For the identified assessments, internal consistency has been reported as follows.

- For the STAR Reading assessment (Renaissance Learning, 2015), overall split-half consistency for Grade 2 has been reported as .95. Reliability for the subtest (“literary text”) will be established on our sample.

- The iReady Reading assessment for second grade reading has marginal reliability for “comprehension: literature” subtest of .84 and test-retest reliability for overall scores (all domains, individual domains not available) of .85.
- For the AimsWeb Plus assessment (NCS Pearson, 2017), internal consistency (Cronbach’s alpha) for the subtest (“reading comprehension”) has been reported as .86.
- For the International Reading Level Assessment (American Reading Company, 2015), the entire assessment will be used. Internal consistency cannot be reported for groups of items given the small numbers of students responding to any specific group of items (Griswold & Bunch, 2014).
- No internal consistency has been reported for the EL module assessments for reading foundations (EL Education, 2016).

Significant improvement is expected in at least STAR Reading, subtest “literary text,” or iReady Reading, subtest “comprehension: literature,” or AimsWeb Plus, subtest “reading comprehension.”

Study B only: At BPS, MAP Fluency/Reading is used as the second grade literacy assessment. MAP Fluency/Reading has been reported to have a marginal reliability of .57 in 2nd grade for language comprehension (foundational skills).

Domain	Name of Instrument	Subtest(s) of instrument used, if any	Timing of measurements	Baseline measure	Variable construction
Computational Thinking	TechCheck	n/a	Right after completion of curriculum (Group 1, 2)	Score before start of curriculum (fall/winter)	Raw score
Coding Knowledge	Coding Stages Assessment	n/a	Right after completion of curriculum (Group 1, 2)	Score before start of curriculum (fall/winter)	Raw score
Comprehension	STAR Reading	Literary text	Spring/summer after curriculum (Group 1, 2)	Score before start of curriculum (fall/winter)	Raw domain score
Comprehension	iReady	Comprehension: literature	Spring/summer after curriculum (Group 1, 2)	Score before start of curriculum (fall/winter)	Domain scale score
Comprehension	AimsWeb Plus	Reading Comprehension	Spring/summer after curriculum	Score before start of	Raw domain score

			(Group 1, 2)	curriculum (fall/winter)	
Reading	International Reading Level Assessment	n/a	Spring/summer after curriculum (Group 1, 2)	Score before start of curriculum (fall/winter)	Scale score
Reading	Fastbridge Reading Screener	n/a	Spring/summer after curriculum (Group 1, 2)	Score before start of curriculum (fall/winter)	Scale score
Comprehension / Reading fluency	MAP Fluency	Comprehension; Reading fluency	Spring/summer after curriculum (Group 1, 2)	Score before start of curriculum (fall/winter)	Raw subtest score; percent correct; pass/fail ²

3.1.6. Attrition (RCTs only)

Both Studies: We will examine attrition of clusters (schools), teachers, and students. Teacher attrition is considered because it affects student attrition. Two types of attrition will be tracked in this study: overall attrition, or the percentage of randomly assigned units in the whole sample for which the researchers do not observe outcome data; and differential attrition, or the absolute value of the percentage point difference between attrition rates for the intervention group and the comparison group. In this section, we describe potential sources of attrition at each level, as well as how attrition rates will be calculated.

Cluster-level attrition: While attrition of entire schools is unlikely, potential reasons for an entire school to be lost from the trial could be a change in leadership, a decision by a school administrator to withdraw due to concerns with implementation or testing, or very high teacher turnover at the school during the implementation period, which could lead the school to withdraw from the study. Of these reasons, the most likely reasons involve school-level leadership.

Attrition at the cluster (school) level will be calculated as the proportion of schools that were randomly assigned to the treatment condition or to the control condition (i.e., the baseline

² MAP Fluency is an adaptive test. The student may receive different language comprehension assessments (listening comprehension, picture vocabulary) based on testing pathways. If the child is working within discrete foundational literacy skills the assessment measures Listening Comprehension, Picture Vocabulary, Phonological Awareness, Phonics and Sentence-level Fluency. If the child has progressed into oral reading it assesses Oral Reading Fluency, including words per minute, accuracy and literal comprehension. At the second-grade level, most BPS students start with the oral fluency test. Approximately 30% of children pass this test and continue to the oral reading section. The rest of the children do not pass the sentence fluency and are re-routed into the foundational skills subtests. We plan to use the sentence fluency threshold (pass/fail) as an outcome of interest in addition to the comprehension scores, since the comprehension assessments vary based on the pass/fail outcome.

sample) that do not contribute data to the analytic sample for an outcome. We will provide information on the numbers of treatment and control schools randomized and the number with data for each outcome. This sample size information will enable overall and differential attrition of schools to be adequately assessed.

Teacher-level attrition: While teacher turnover may be an issue in either study, especially over the summer in Rhode Island, the turnover over the relevant period of 3 months during fall/winter/early spring is typically low. It is more likely that teacher attrition from the study may occur for other reasons, including failure of following through on data submission (e.g., lesson logs, surveys).

Attrition at the teacher level will be calculated as the proportion of teachers that were randomly assigned to the treatment condition or to the control condition (i.e., the baseline sample) that do not contribute data to the analytic sample for an outcome. We will provide information on the numbers of teachers in treatment and control schools randomized and the number with data for each outcome. This sample size information will enable overall and differential attrition of teachers to be adequately assessed.

Student-level attrition: Students will be lost from the analytic sample if they leave a treatment or control school during the 3-months implementation period. It is more likely that student attrition from the study may occur for other reasons, particularly incomplete data submission (e.g., standardized test data).

Attrition at the student level will be calculated as the proportion of students that were randomly assigned to the treatment condition or to the control condition (i.e., the baseline sample) that do not contribute data to the analytic sample for an outcome. We will provide information on the numbers of students in treatment and control schools randomized and the number with data for each outcome. This sample size information will enable overall and differential attrition of teachers to be adequately assessed.

We will deploy three primary strategies to minimize cluster-, teacher-, and student-level attrition:

- ***Embedded Coach as liaison:*** A key component of the program design is embedded coaching. In addition to supporting teachers, the coach assigned to a school will act as a primary point of contact for the Intervention and Research and Evaluation teams, relaying any concerns with curriculum implementation and/or evaluation, assisting with arrangements for testing, and acting as an “on the ground” liaison with school administrators and teachers. The Research and Evaluation team also will closely monitor lesson logs, survey responses, and testing participation rates during the study period and will identify schools where a substantial number of teachers are not submitting data, individual teachers with uneven response rates to surveys, logs, and assessments, and individual classrooms with students with incomplete assessments. This information will

be shared with the embedded coach, who will reach out to the school administrator and teachers directly to check in on implementation and offer support as needed. This front-line support will enable the curriculum and evaluation teams to be quickly informed about issues at the school, classroom, and teacher levels and make adjustments/intervene as needed.

- ***Frequent communications with district and school administrators:*** The Intervention team will maintain frequent communications with district and school administrators and teachers prior to and throughout the study period.
 - Administrators will be briefed immediately prior to the start of the school year; in addition to briefing about the logistics of the study, the administrators will learn about the CAL curriculum and its projected benefits. Follow-up communications will occur with administrators prior to and following teacher training; administrators at schools participating in the curriculum will receive copies of training materials and notification of teachers' attendance at, and completion of, the training. Administrators of schools implementing the CAL curriculum will be contacted by the Research and Evaluation team prior to the start of curriculum implementation to coordinate testing periods. All administrators will be contacted once again by the Research and Evaluation team at the midpoint of curriculum implementation to check for any concerns and to coordinate testing periods.
 - The Intervention team will maintain frequent communications with teachers prior to and throughout the study period. Email communications with teachers will begin prior to training. Following the training period, the Intervention team will send email communications to teachers implementing the CAL curriculum, providing supplemental information on the curriculum, instructional ideas, and reminders about the resources available to support instruction, in addition to scheduling testing. The Intervention team will also check in with teachers at the midpoint of curriculum implementation to check for any concerns and to schedule testing.
- ***Site-level monitoring during testing periods:*** The Research and Evaluation team will maintain close contact with research assistants conducting testing. Research assistants will maintain an online implementation log and will be directed to identify any concerns or incidents that take place during testing periods. During the testing periods, the Research and Evaluation team will monitor research assistant logs daily and will take action (e.g., contact school administrator and/or teacher) if there are issues with access to students or other concerns.
- ***Flexibility in test scheduling:*** Scheduling will be coordinated with school administrators to avoid scheduling conflicts at the school related to other priorities (e.g., standardized testing). Recognizing that all students may not be in attendance on any given day, the

Research and Evaluation team will schedule additional time periods for each classroom to test students that may have been absent during the initial testing day.

3.1.7. Statistical analysis of impacts

3.1.7.1. Contrasts

The contrasts can be found in the contrast tool (*2000917_Early52_contrast_tool.xlsm*).

3.1.7.2. Strategy for dealing with multiple comparisons

There will be one domain that has multiple outcomes: the comprehension domain. The evaluator will use the Benjamini-Hochberg method to adjust for multiple comparisons in each domain.

3.1.7.3. Model specifications

The model is a basic cluster assignment without blocking, with two levels: the student level and the school level.

Level-1: Student Level

$$Y_{ijk} = \beta_{0j} + \beta_{1j}(Y_{ij}^*) + L_i + \sum_{m=1}^M \beta_{2.mj} X_{mij} + \varepsilon_{ij}$$

Level-2: Cluster (School) Level

$$\beta_{0j} = \gamma_{00} + \gamma_{01}(T_j) + \sum_{q=1}^Q \gamma_{02.q} W_{qj} + \mu_{0j}$$

$$\beta_{1j} = \gamma_{10}$$

$$\beta_{2.mj} = \gamma_{2.m0}$$

Where,

Y_{ij} = the outcome score for the i^{th} student in the j^{th} school.

β_{0j} = the intercept for school j .

β_{1j} = the effect of pretest in school j .

Y_{ij}^* = a pre-test measure for the i^{th} student in the j^{th} school.

$\beta_{2.mj}$ = the effects of student covariates in school j .

X_{mij} = the m^{th} of M additional covariates for student i in school j .

L_i = 1 if teacher of student i is a TechLeader, = 0 if school j is not a TechLeader.

ε_{ij} = a residual error term for student i in school j .

γ_{00} = the mean intercept

γ_{01} = estimated treatment impact

T_j = 1 if school j is assigned to treatment (CAL), and = 0 if school j is assigned to comparison.

$\gamma_{02.q}$ = the effect of school-level covariate (percent of students receiving free/reduced-price lunch);

W_{qj} = the q^{th} of Q covariates for school j .

μ_{0j} = random intercept term – deviation of cluster j 's mean from the grand mean, conditional on covariates; assumed to be normally distributed with mean 0 and variance τ_{00}^2 .

γ_{10} = mean effect of pretest

$\gamma_{2.m0}$ = mean effect of student covariate m .

The parameter estimate γ_{01} provides a covariate-adjusted estimate of the impact of the CAL curriculum. The hypothesis test for γ_{01} will determine whether or not the intervention has a statistically significant impact on the given outcome. A standardized effect size will be calculated by dividing the impact estimate (γ_{01}) by the pooled standard deviation derived from a) the published population standard deviations for comprehension outcomes and b) the unadjusted sample standard deviations for the outcome in the intervention and comparison groups. We will estimate this HLM model for all six student outcomes.

3.1.7.4. Subgroups

N/A

3.1.7.5. Decision rules for inclusion/exclusion of covariates

The selection of covariates is theory-driven. We will be including covariates that have in previous research been found to affect the student outcomes measured in this intervention, and these will not be eliminated, even if they turn out to be not significant in our sample. Students' baseline measures will always be included in the model.

3.1.7.6. Treatment of missing data

Missing outcome data will not be imputed. Cases with missing outcome data will be deleted. If outcome data is available for a subset of the measures (e.g., just for comprehension measures), the case will be included in the analytical sample for that outcome but dropped for those where outcome data is not available. Missing values for pre-tests and for covariates will be imputed using dummy variables.

3.1.7.7. Calculation of effect size

We will use Hedges' g as our effect size index. For the standardized student outcomes (comprehension standardized assessments) we will use population standard deviation (SD) values, if available from the publisher. If these are not available, we will use an estimate from the data, with pooled SD. For the CSA and TechCheck, we will use an estimate from the data, with pooled SD.

3.1.8. Baseline equivalence testing – QEDs, RCTs with high attrition, cluster RCTs with potential bias from entering individuals

Both studies use random assignment, so we assume that any differences between treatment and comparison group students both on observable and on unobservable variables that exist at baseline occur purely by chance. In the event of high attrition, we will assess the equivalence of the treatment and control students at baseline for each analytic sample. If attrition is high, the analytic sample will be defined as students without a missing outcome and without missing data for baseline (pretest) measures. Analytic samples for each outcome may vary slightly, given differences in missing data; therefore, baseline equivalence will be assessed for each analytic sample. We will report the mean and standard deviation of each baseline measure, along with the sample size for each group at baseline. Because we will be using individual level data from the same students at baseline and post-treatment, representativeness of individuals of clusters is not a concern.

To assess baseline equivalence for all pre-test measures (TechCheck, Coding Stages Assessment, literacy comprehension), we will use a two-level hierarchical linear model that reflects the features of the design (i.e., random cluster assignment without blocking). Specifically, we will adapt the model used to measure impacts (see section 3.1.7.3, model specifications), but moving the baseline measures to the left of the equation and removing all covariates. The parameter estimate for the treatment variable (γ_{01}) will provide an estimate of the magnitude of the baseline mean difference between the treatment and comparison students in the scale of the baseline measure.

We will calculate the standardized baseline difference (Hedges' g) by dividing the parameter estimate (i.e., γ_{01}) by the pooled standard deviation derived from the unadjusted sample standard deviations for the intervention and comparison groups.

Level-1: Student Level

$$BaseVar_{ij} = \beta_{0j} + \varepsilon_{ij}$$

Level-2: Cluster (School) Level

$$\beta_{0j} = \gamma_{00} + \gamma_{01}(T_j) + \mu_{0j}$$

Where,

$BaseVar_{ij}$ = the baseline measure for the i^{th} student in the j^{th} school

β_{0j} = the intercept for school j .

γ_{00} = the unadjusted mean baseline value for schools in the comparison group

γ_{01} = the difference in means between schools in the treatment and in the comparison groups.

T_j = 1 if school j is assigned to treatment (CAL), and = 0 if school j is assigned to comparison.

μ_{0j} = random intercept term for school j

ε_{ij} = residual error term for student i in school j

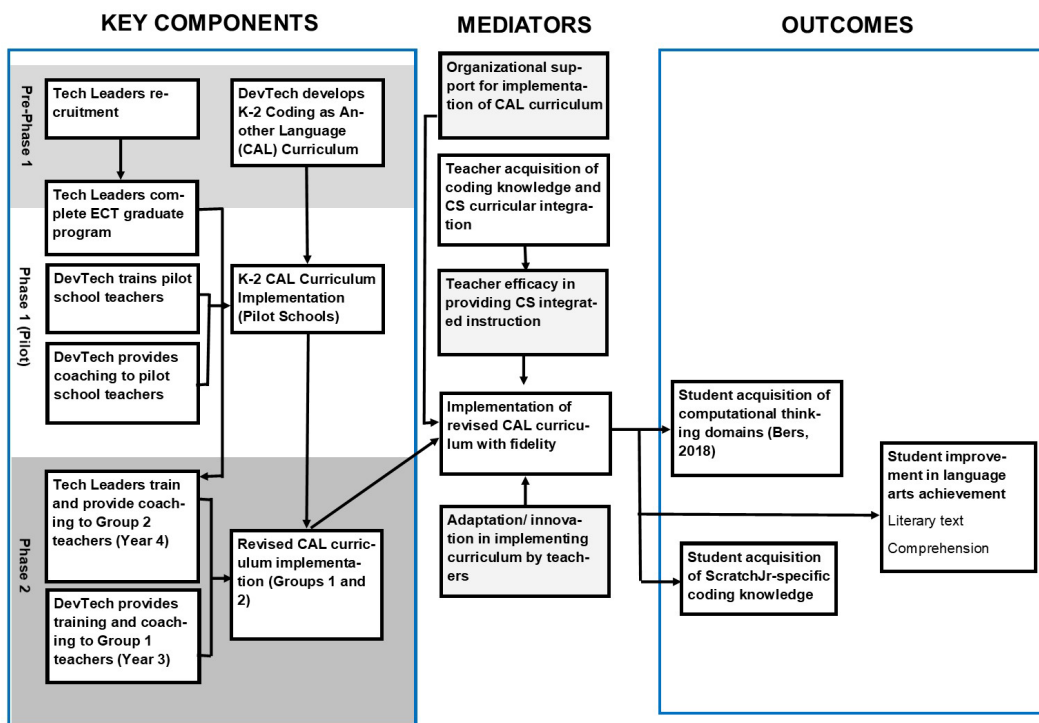
4. Implementation Evaluation

4.1. Implementation Study of the EIR-funded Intervention

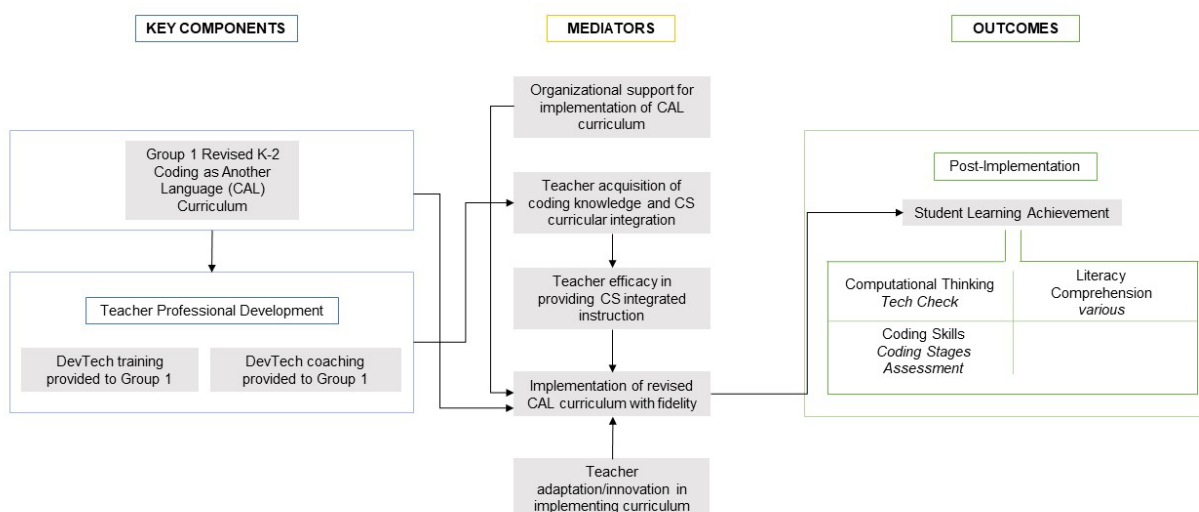
4.1.1. Logic model for the intervention

We created two logic models. The first one gives an overview of the entire project, which includes a pilot scale-up study. Part of the pilot scale-up study is the training of Tech Leaders, who will be implementing the intervention in Year 4 of the project. However, the scale-up study is not part of the evaluation. For the purpose of the evaluation, the second logic model has been created, which only contains the key components, mediators and outcomes associated with the impact and implementation study.

Logic model for the entire project:



Logic Model for Impact and Implementation Study:



To improve student learning outcomes, teachers need 1) a developmentally appropriate research-based integrated CS curriculum, combined with 2) comprehensive, multiple, and ongoing forms of professional development to support implementation fidelity. Additionally, there is a need to develop capabilities of Tech Leaders, which have a crucial role to support the integration and teaching of CS in K-2.

The first key component for the project is to develop a comprehensive research-based integrated CS curriculum (CAL curriculum) for K-2 and distribute a copy of this curriculum to teachers. The curriculum will meet the following principles: 1) flexible project-based units that introduce coding and computational thinking in a playful, developmentally appropriate way by integrating powerful ideas of computer science with literacy skills; 2) strengthened social-emotional aspects by not only focusing on the cognitive dimension of computer science, such as problem solving, but also habits of mind such as perseverance; 3) the format and content of activities explicitly designed to be attractive to girls and marginalized populations (Fisher & Margolis, 2002; Richmond, 2000; Rosser, 1990; Sadler, Coyle, & Schwartz, 2000; Tobin, Roth, & Zimmerman, 2001), as well as meet the needs of gifted or special education students by being adaptable and personalized (O'Connor, 2000).

A draft version of the curriculum was based on previous work by DevTech, a scope and sequence of K-2 integrated curriculum has been developed in consultation with experts to ensure that the curriculum can be used by every student at the sites and be fully integrated. This draft curriculum was then piloted in four schools, which were selected to represent a range of experiences and backgrounds to ensure the curriculum and supports meet the needs of a diverse

workforce. Schools that participate in the piloting will not be placed in the pool to be assigned to the implementation study. Based on this pilot, the draft was revised, and the final curriculum will be developed. The final (revised) curriculum will be implemented during Year 1 in 5 schools in Rhode Island and in approximately 10 Boston Public Schools using a delayed treatment design.

The second key component for the project is to provide *DevTech training* to teachers in Group 1 schools. The teacher training consists of four hours of online training to understand the ScratchJr app and the developmental underpinnings, the scope and sequence of the CAL curriculum, the points of alignment and integration with other disciplinary content and skills, the adaptations that can be made to better suit the population needs, and the embedded student's assessments within the curriculum.

The third key component, *DevTech coaching*, provides job-embedded professional development to Group 1 teachers that builds on the training. It involves 1) embedded on-site coaching and 2) individualized online virtual coaching sessions. Teachers also have the option to participate in the ScratchJr Connect on-line network. Teachers will receive a regular email to alert them of new learning opportunities and video case studies aligned with the curriculum and will be able to sign up for individualized targeted on-line or on-site tutoring sessions.

There are several **mediators** or **short-term outcomes** that will affect the ultimate long-term learning outcomes for students. Two key mediators will be measured: implementation with fidelity of the revised CAL curriculum, teacher efficacy in providing computer science integrated instruction (Isbell & Szabo, 2015; Tschannen-Moran, Hoy & Hoy, 1998), and teacher acquisition of coding knowledge (Guskey, 2000). Other variables (shaded on the model) will not be measured in the impact study but are acknowledged as mediators: organizational support for implementation of the computer science integrated curriculum (Abdal-Haqq, 1996; Guskey, 2000; Ingvarson & MacKenzie, 1988); and teacher adaptation and innovation in implementing the curriculum (Hord & Roussin, 2013; Quinn & Kim, 2017).

The ultimate long-term outcome is for students to have positive learning achievement, especially in the areas of computational skills; ScratchJr-specific coding knowledge; and in literacy the domain of comprehension.

4.1.2. Research questions for evaluation of implementation

1. To what extent were the key components of the intervention implemented with fidelity?
2. What was the amount of variation in implementation fidelity?
3. What was the relationship of fidelity of implementation to intermediate outcomes associated with changes in faculty, coaches, counselors, or other individuals implementing the intervention?

4.1.3. Measuring fidelity of implementation

The first key component for the project is to develop a comprehensive research-based integrated CS curriculum (CAL curriculum) for K-2 and distribute an online copy of this curriculum to teachers (*pre-curriculum*).

The second key component for the project is to provide *DevTech training* to teachers in Group 1 schools. The teacher training consists of full-day seminars to understand the developmental underpinnings, the scope and sequence of the CAL curriculum, the points of alignment and integration with other disciplinary content and skills, the adaptations that can be made to better suit the population needs, and the embedded student's assessments within the curriculum.

The third key component, *DevTech coaching*, provides job-embedded professional development to Group 1 teachers that builds on the training. It involves 1) embedded on-site coaching and 2) individualized online virtual coaching sessions. Teachers also have the option to participate in the ScratchJr Connect on-line network. Teachers will receive a weekly email to alert them of new learning opportunities and video case studies aligned with the curriculum and will be able to sign up for individualized targeted on-line and on-site tutoring sessions.

Key component	Indicator	Indicator description	Unit	Threshold at unit level	Threshold next level	Threshold at sample level
Revised CAL curriculum	Dissemination of curriculum	Any entry in teacher's lesson log	Teacher	Teacher level threshold: 2 2 = have it at the start of teaching	School level threshold: 3 2 = 75-89% teachers with score of "2"	Sample level threshold: 3 3 = 75-89% schools with score of "2"
DevTech training	Group training participation	Attendance sheets collected at training by training facilitator	Teacher	Teacher level threshold: 3 3 = attended 75% to 89% of training	School level threshold: 3 3 = 75-89% teachers with score of "3" or more	Sample level threshold: 3 3 = 75-89% schools with score of "3" or more
	Group training topic content	Topic list checked by	Sample	N/A	N/A	Sample level threshold: 3

Key component	Indicator	Indicator description	Unit	Threshold at unit level	Threshold next level	Threshold at sample level
		external observer				3 = covered 75% to 89% of topics
All indicators training				Teacher level threshold: 3	School level threshold: 3 3 = 75-89% teachers with score of “3” or more	Sample level: Range: 0-8 Threshold for fidelity = score of 6
Coaching	Embedded onsite coaching	Availability of onsite coaching	Teacher	Teacher threshold level: 2 2 = teacher received a response to all requests made		
		Satisfaction with onsite coaching	Teacher	Teacher threshold level: 1 1 = (“meets expectations”)		
	Virtual coaching	Availability of virtual coaching	Teacher	Teacher threshold level: 2 2 = teacher received a response to all requests made		
		Satisfaction with virtual coaching	Teacher	Teacher threshold level: 1 1 = (“meets expectations”)		
All indicators coaching				Adequate implementation with score of at least 3 (if one of the coachings is N/A) or 6. If no training is	School level threshold: 3 3 = 75-89% teachers with	Sample level threshold: 3

Key component	Indicator	Indicator description	Unit	Threshold at unit level	Threshold next level	Threshold at sample level
				accessed, the threshold is N/A.	score of “3” or “6” or more (excluding N/As)	3 = 75-89% schools with score of “3”

4.1.4. Data collection plan

Note: Despite the complications caused by the COVID-19 pandemic, the schedule for the study remains unchanged, with the impact study planned to take place in 2021/2022. Note that the dates differ for the two sites: Data collection will start in Rhode Island in summer 2021 and in BPS in fall 2021.

Indicator	Data source	When and how collected
Dissemination of curriculum	Lesson logs	Any entry in teacher’s lesson log
Group training participation	Attendance sheets	Collected at PD training (Summer/Fall 2021) by evaluator
Group training content	Checklist of topics	Filled in by evaluator at PD training or following the training using a recording of the PD training event (Summer/Fall 2021)
Coaching (Tier 1 embedded onsite coaching; Tier 2: virtual 1-on-1 coaching): Availability and satisfaction with coaching	Teacher survey	Teachers report on availability and satisfaction with coaching in survey; evaluator downloads survey results after implementation (January/May 2022)

4.1.5. Fidelity reporting plan

At or prior to the end of the grant period, the evaluation team will report fidelity results for one year of implementation. (*This project has prior OII approval of reporting fidelity for only one*

year of implementation.) Using the reporting template below, the evaluation team will report fidelity by key component of the intervention (i.e., CAL curriculum/ “pre-curriculum”, DevTech training, and DevTech three-tiered coaching). Fidelity scores will be reported by indicator for the entire implementation sample. Finally, based on the specification of thresholds for determining fidelity, the team shall report whether each key component was implemented with fidelity at the sample level.

Key Component x (of 3) –[Key component name]. Fidelity Matrix and Fidelity Results Reporting Table

Indicators	Definition	Unit of implementation	Data Source(s)	Data Collection (who, when)	Score for levels of implementation at unit level	Threshold for adequate implementation at unit level	Roll-up to next higher level if needed (score and threshold): Indicate level	Roll-up to next higher level if needed (score and threshold): Indicate level	Roll-up to sample level (score and threshold for adequate implementation at sample level)	Expected sample for fidelity measure (n = # units in which the intervention is being implemented)	Expected years of fidelity measurement
Key component score											
					Threshold					# of Units Measured (of X units served)	Year of Measurement
Fidelity Results					Achieved Score at Sample Level:						
					Met Threshold Implemented with Fidelity (Yes, No, N/A)						

4.1.6. Plan for providing performance feedback to inform project design (Early-phase grants only)

The pilot study took place during the 2020-2021 school year in four schools: Belgrade-Brooklyn-Elrosani Elementary in Belgrade, Minnesota; Terry Elementary in Little Rock, Arkansas; Lafayette Elementary in San Francisco, California, and Sanchez Elementary in San Francisco,

California. The pilot study served two core functions for this project. First, the pilot study tested and provided feedback to improve the key components – the CAL curriculum for K-2 and professional development that includes both training and coaching for teachers and Tech Leaders. Secondly, the pilot study allowed the research team to conduct a small-scale test of the research methods and procedures that will be used in the full-scale study to ensure the feasibility of the research approach.

For testing of the research methods and procedures, all data collection methods and procedures that will be used in the implementation and impact studies were used in the pilot study. During pilot study planning and implementation, the evaluation team met weekly to review actions and resolve issues, documenting decisions in an ongoing log. The evaluation team is currently reviewing the log and will recommend changes to the impact and implementation study plans. The pilot study included additional data collection to test and provide feedback to improve the key components (i.e., curriculum and professional development). Focus groups with teachers were held immediately following training, and interviews were conducted with teachers at the midpoint of and following curriculum implementation. School administrators (i.e., principals for each school) were interviewed following curriculum implementation. These, along with a teacher survey conducted by the evaluator, collected critical input addressing the following key questions:

- What do teachers/administrators find useful? Not useful?
- What aspects of the curriculum were challenging to implement?
- What supports/resources are essential for successful implementation?
- What are the implementation barriers?

A supplemental analysis of children's resulting ScratchJr. projects also provided insight into children's coding skills and creativity. The pilot study provided an opportunity to assess the degree to which the key mediators—implementation with fidelity of the revised CAL curriculum, teacher efficacy in providing computer science integrated instruction, and teacher acquisition of coding knowledge—affected student outcomes. The first mediator was assessed through the fidelity of implementation study (see section 4.1.3) supplemented by lesson logs maintained by participating teachers, while the second and third mediator were assessed through Computing Beliefs and Computing Self-Efficacy scales (Rich 2017) and the Coding Stages Assessment (Bers 2019) respectively. We found that FOI for the curriculum was high across the sample (meeting the threshold for all four indicators), so that there was not enough variation to use the variable as a mediator. At the time of writing, the student outcomes and surveys are still being analyzed to look at potential mediating effects of self-efficacy and teacher knowledge of coding.

The pilot study also allowed the Research and Evaluation team to track the percentage of teachers utilizing the coaching outside of professional development. We found that 6 out of 10 teachers used virtual and/or onsite coaching. This is a higher proportion than anticipated. The quality of coaching met or exceeded teachers' expectations, so that the researchers decided to maintain the same model for the impact study. The Research and Evaluation team will prepare for the implementation team a detailed report on pilot study findings about key components and facilitate a discussion to review the findings in detail. Minutes of this discussion will document specific recommendations for improvements to the key components.

5. Alignment of Evaluation Samples with Units Receiving the Intervention

The intervention is implemented in 15 schools at the two sites, in grades K-2. However, because participation of teachers is voluntary, at a given school, not all classrooms of the same grade necessarily participate, although teacher participation will be identified prior to randomization. The impact sample comprises 31 schools (15 schools implementing the intervention, 16 schools in the control group). All students in participating Grade 2 classrooms in these schools comprise the impact sample. Fidelity of implementation will be assessed in all schools in the treatment group.

GrantID:				
		A	B	C
	Grant Year	Implementing sample (All schools/ units where EIR-funded intervention is being implemented at any time during the grant period, NOT INCLUDING pilot schools)	Impact sample (including any members of the impact sample during the impact study period)	Fidelity sample
Total number of schools by year (or teachers for teacher/classroom assignment)	1			
	2	15	15	15
	3	16	16	16
	4			
	5			
Grade span by year	1			
	2	Gr. 2	Gr. 2	Gr. 2
	3	Gr. 2	Gr. 2	Gr. 2
	4			
	5			
Types of students included/excluded (e.g., intervention serves all students within	1			
	2	All students included	All students included	All students included
	3	All students included	All students included	All students included
	4			

GrantID:				
		A	B	C
	Grant Year	Implementing sample (All schools/ units where EIR-funded intervention is being implemented at any time during the grant period, NOT INCLUDING pilot schools)	Impact sample (including any members of the impact sample during the impact study period)	Fidelity sample
relevant grades) by year Compare columns A & B: Reason for/basis if impact sample (B) does not include all units in implementation sample (A) in each year	5			
	1			
	2			
	3			
	4			
	5			
Compare columns A & C: Reason for/basis if fidelity sample (C) does not include all units in implementation sample (A) in each year	1			
	2			
	3			
	4			
	5			

6. References

- Abdal-Haqq, I. (1996). Making Time for Teacher Professional Development. *ERIC Digest*.
- Bers, M. U. (2018). *Coding as a Playground* (1st ed.). Routledge.
- Bers, M. U. (2019). Coding as another language: A pedagogical approach for teaching computer science in early childhood. *Journal of Computers in Education*, 6(4), 499–528.
<https://doi.org/10.1007/s40692-019-00147-3>
- Guskey, T. R. (2013). Does It Make a Difference? Evaluating Professional Development. *ASCD*, 13.
- Hord, S. M., & Roussin, J. L. (2013). *Implementing Change Through Learning: Concerns-Based Concepts, Tools, and Strategies for Guiding Change*. SAGE Publications, Ltf.
- Ingvarson, L., & MacKenzie, D. (1998). Factors Affecting the Impact of Inservice Courses for Teachers: Implications for Policy. *ERIC Digest*.
- International Technology Education Association. (2007). *Standards for Technological Literacy*.
- Isbell, L., & Szabo, S. (2015). Assessment: Teacher Efficacy and Response to Intervention. *The Delta Kappa Gamma Bulletin*, 81(No.2).
- K12 Computer Science Framework. (2016). Retrieved from <http://www.k12cs.org>.
- Margolis, J., & Fisher, A. (2002). *Unlocking the clubhouse: Women in computing*. MIT press.
- National Governors Association Center for Best Practices & Council of Chief State School Officers. (2010). *Reacher Higher: The Common Core State Standards Validation Committee*.
- O'Connor, B. (2000). Using the design process to enable primary aged children with severe emotional and behavioral difficulties (EBD) to communicate more effectively. *The Journal of Design and Technology Education*., 5 (3), 197-201.

- Quinn, D. M., & Kim, J. S. (2017). Scaffolding Fidelity and Adaptation in Educational Program Implementation: Experimental Evidence From a Literacy Intervention. *American Educational Research Journal*. <https://journals-sagepub-com.ezproxy.library.tufts.edu/doi/full/10.3102/0002831217717692>
- Relkin, E., de Ruiter, L., & Bers, M. U. (2020). TechCheck: Development and Validation of an Unplugged Assessment of Computational Thinking in Early Childhood Education. *Journal of Science Education and Technology*, 29(4), 482–498. <https://doi.org/10.1007/s10956-020-09831-x>
- Renaissance Learning. (2015). *STAR Reading Technical Manual*. Renaissance Learning.
- Rich, P.J., Jones, B., Belikov, O. Yoshikawa, E., & Perkins, M. (2017). Computing and Engineering in Elementary School: The Effect of Year-long Training on Elementary Teacher Self-Efficacy and Beliefs about Teaching Computing and Engineering. *International Journal of Computer Science Education in Schools* 1(1), 1-20.
- Richmond, G. (2000). Exploring the complexities of group work in science class: A cautionary tale of voice and equitable access to resources for learning. *Journal of Women and Minorities in Science and Engineering*, 6(4), 295-311.
- Rosser, S.V. (1990). *Female-friendly science: Applying women's studies methods and theories to attract students*. London: Pergamon.
- Rutterford, C., Copas, A., & Eldridge, S. (2015). Methods for sample size determination in cluster randomized trials. *International Journal of Epidemiology*, 44(3), 1051-1067.
- Sadler, P. M., Coyle, H. P., & Schwartz, M. (2000). Engineering competitions in the middle school classroom: Key elements in developing effective design challenges. *The Journal of the Learning Sciences*, 9(3), 299-327.

Tobin, K., Roth, W.M., & Zimmermann, A. (2001). Learning to teach science in urban schools.

Journal of Research in Science Teaching, 38(8), 941-964.

Tschannen-Moran, M., Hoy Woolfolk, A., & Hoy, W. K. (1998). Teacher Efficacy: Its Meaning and Measure. *Review of Educational Research*.

<http://journals.sagepub.com/doi/10.3102/00346543068002202>

Virginia Department of Education. (2017, November). *Computer Science Standards of Learning Resources*. Virginia Department of Education.

http://www.doe.virginia.gov/testing/sol/standards_docs/computer-science/index.shtml