A Summary of the BURST®: Reading Efficacy Trial

(Carried Out Under IES Award # R305A120811**)

Report Authors:
Brian Rowan (Project Director), Ben B. Hansen (Participating Investigator),
Mark White (Research Fellow), Timothy Lycurgus (Graduate Assistant),
Lesli J. Scott (Senior Survey Director)

Institute for Social Research
Survey Research Center
University of Michigan

March 2019

**Data collection and analysis for this research were funded by a grant from the Institute for Education Sciences (Award #: R305A120811) to the University of Michigan and Amplify, Inc. We thank: Sandra Pappas, Allison York, and Kathleen Richards of Amplify, Inc. for useful assistance in data collection; Dolorence Okullo of the University of Michigan for assistance in data preparation; and Stephen G. Schilling for assistance in the analysis of teacher survey data. The authors of this paper are solely responsible for preparation of data tables and interpretations of results reported here.

Abstract

This report summarizes the results of a cluster-randomized field trial that estimated the effect of *BURST®: Reading* on primary grades students' early literacy achievement. BURST is a widely adopted supplemental reading program designed for use with students struggling to acquire early literacy skills and is meant to provide supplemental instruction to these students outside the regular reading program. The program uses an "assess, group, instruct" format in which schools identify struggling readers using the *DIBELS Next* assessment, then use a proprietary algorithm to place identified students into reasonably homogenous skill groups on the basis of DIBELS results, and then provide targeted instruction to these groups using BURST curriculum and lesson materials. Ober the four-year period AY 2013-2014 to AY 2016-2017, the University of Michigan (in co-operation with Amplify, Inc.) carried out a study in 52 high-poverty schools serving grades K-3 located in 9 states in different geographic areas of the United States during the period AY 2013-2014 to AY 2016-2017. The study randomly assigned 27 schools to the BURST treatment group and 25 to a control group that was provided free access to the *DIBELS Next* assessment for use in a regular universal screening process. More than 29,000 students enrolled in grades K-3 at treatment and control schools participated in the study contributing about 1.8 observations per student. Data analysis showed no evidence of differential attrition in the study groups, there was strong evidence of baseline equivalence of the treatment and control samples in the study, and cross-over from one experimental condition to the other was minimal and similar across treatment and control groups. The study found that schools assigned to the BURST treatment group offered BURST instruction to both struggling and non-struggling readers and that the average struggling reader received about 40 hours of BURST instruction in a given a year such that (over a four year period) the average struggling reader in a BURST school could be expected to accumulate between 120 and 140 hours of BURST instruction. Overall rates of provision of BURST instruction in study schools was found to be similar to rates of provision of BURST instruction in schools with similar demographic characteristics that had purchased and were using the BURST program outside the efficacy trial in AY2016-2017 but less than the amount of instruction recommended by the program developer. Using the Star Early Literacy assessment as the primary outcome, and after adjusting this outcome for several pre-treatment covariates using a Peters-Belson type strategy, the study estimated sample average treatment effects on students' early literacy learning using permutation tests that took into account the various forms of clustering in the experimental design and that controlled statistical significance tests for family wise error rates due to multiple comparisons. The results of these statistical tests showed that the BURST program did not have statistically significant effects on the early reading achievement of all students who attended BURST schools, did not have statistically significant positive effects on the early reading achievement of struggling readers who attended BURST schools, did not have statistically significant positive effects on the early reading achievement of students who attended BURST schools for three or more years consecutively, and did not have statistically significant positive effects on the early reading achievement of students who attended BURST schools that had a predicted probability greater than 1% of complying (versus not complying) with treatment assignment. There was only slight evidence of school-to-school variability in program effects, with schools implementing BURST instruction at higher rates tending to have slightly larger positive effects on students' early reading achievement than schools implementing BURST instruction at lower rates. These differences were very small, however, and as such, were not assessed for statistical significance. In all, the study's results are best summarized as follows: In a sample of 52 schools located in school districts that are smaller than the average U.S. school district, in communities that are more disadvantaged than average, and serving higher percentages of lower achieving students than average, the added benefit of using BURST for supplemental reading instruction under routine conditions of implementation was found to be negligible compared to engaging in universal screening with DIBELS and conducting supplemental reading instruction under business as usual conditions.

1 | Introduction

Overview

This report summarizes the results of a cluster-randomized field trial that estimated the effect of *BURST®: Reading* on primary grades students' early literacy achievement. BURST is a supplemental reading program intended for use in grades K-6 that was developed and is marketed by Amplify, Inc., a Brooklyn-based vendor of curricula, assessments, and intervention programs for K-12 schools. In 2012, the University of Michigan and Amplify. Inc. received a grant from the Institute for Education Sciences (IES) to conduct a randomized controlled efficacy trial (RCT) of this program. The RCT was conducted over a four-year period beginning in September 2013 with samples of students in kindergarten to third grade who were located in 52 schools, 27 of which were randomly assigned to the BURST treatment condition and 25 of which were randomly assigned to a control condition. In each of the four years of the study, data on program implementation were collected in treatment schools and data on student achievement were collected in both treatment and control schools. This report summarizes the University of Michigan's analyses of data from this study and builds on those analyses to draw some inferences about the effectiveness of *BURST®: Reading* for improving the early literacy achievement of students in grades K-3.

Conflict of Interest Statement

Because the BURST efficacy trial involved direct participation of the program vendor (Amplify, Inc.), the University of Michigan and Amplify took steps to safeguard data collection, analysis, and reporting processes from potential conflicts of interest. These steps are described below as a prelude to this report.

Recruitment of Schools and Assignment to Treatment: A strict division of labor between Amplify and the University of Michigan governed the recruitment and random assignment of schools in the study. The program vendor (Amplify) recruited schools into the study, but schools had to agree to join the study in advance of knowing whether they would be randomly assigned by the University of Michigan to the treatment or control condition. As an incentive to join the study, Amplify offered free company services to schools. Control schools were given a free subscription to Amplify's digital version of *DIBELS Next®* (a formative assessment of students' early literacy skills) along with free access to all of the training and services normally provided to regularly subscribing schools. Treatment schools were given a free subscription to *DIBELS Next®* plus free access to the *Burst®: Reading* program, along with free access to all of the training, materials, and implementation support normally provided to regularly subscribing schools. Once schools were enrolled in the study, University of Michigan researchers randomly assigned schools to treatment or control conditions using procedures described later in this report. After assignment, Amplify provided each school free access to the training and services appropriate to its assigned condition.

Data Collection: Amplify conducted most of the data collection activities for the study, but did so under the supervision of University of Michigan research staff. Using its regular business systems and processes, Amplify gathered data on program implementation in treatment schools during each year of the study. Amplify also worked with schools to conduct the annual achievement testing of students in treatment and control schools. Under the supervision of University of Michigan research staff, and outside its normal business systems, Amplify also worked at school sites to assemble and verify rosters teachers and students at each study school. All of the data collected by Amplify were securely transferred to University of Michigan research staff shortly after collection, and University of Michigan research staff inspected all data for quality control purposes. An important step in this quality control process involved careful matching of student achievement and other data collected by Amplify business systems to student and teacher rosters verified on site. This step assured an accurate accounting of all students eligible for inclusion in the study, and for careful monitoring of patterns of missing student data and student attrition across study years.

Data Analysis: At the end of the four-year study period, only University of Michigan researchers had access to the complete, compiled, and cleaned data used in the analyses reported here. Prior to conducting an outcomes analysis of these data, University of Michigan researchers registered a data analysis plan with the *Registry of Effectiveness and Efficacy Trials* maintained by the Society for Research on Educational Effectiveness (Registry ID 473). The registry of a study plan allows readers of this report to judge the results reported here against a "transparent" and publicly available data analysis plan prepared *prior to* actual outcomes analysis.

Structure of the Report

This report proceeds as follows: In the next section, we briefly describe the *BURST®: Reading* program. In a following section, we list the research questions we set out to address in the current study. In a subsequent section, we describe the schools selected for inclusion into the study, the patterns of attrition that occurred over the course of the study, and a set of baseline comparisons that were used to assess any differences in pre-treatment treatment covariates across treatment and control schools that might be associated with the student achievement outcome on which program effects are estimated. In the section following, we describe our analyses of program implementation data, and in the section following that, our analyses of BURST program effects on student achievement. A final section briefly summarizes the results of the study.

2 | The Burst®: Reading Program

Overview

BURST®: Reading is a supplementary instructional program designed to help elementary schools identify and provide targeted instruction to struggling readers. Developed by Amplify, Inc., the program was first brought to market in 2007. As of 2016, it was in use in 651 schools across the United States.

Program Logic

As a supplementary intervention program, *BURST®: Reading* is designed to identify and provide targeted instruction to students who are struggling to acquire such early literacy skills such as phonological knowledge, knowledge of the alphabetic principle, and oral fluency in reading. The design for implementation in schools is as follows:

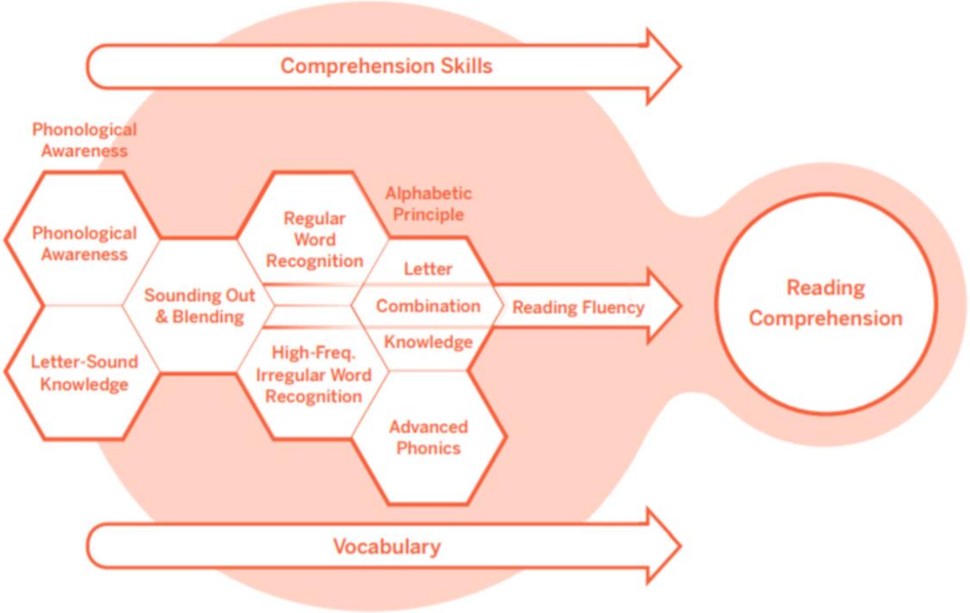
- At the beginning (and again at the middle) of the year, *all* students in participating grades at a school are assessed with the Dynamic Indicators of Early Literacy Skills (*DIBELS Next®*) assessment.
- Based on this universal screening process, schools identify students whose DIBELS assessment scores are below or well below DIBELS composite score benchmarks for expected performance for a given grade and time of year.
- Schools then choose which of these identified students will be assigned to BURST instruction.
- Once schools identify students, Amplify applies a proprietary algorithm that places students into relatively homogeneous small groups for indicated BURST instruction. Indicated instruction involves teacher-led provision of a set of sequenced lessons (hereafter called BURST cycles) to small groups of students who have similar *DIBELS Next®* assessment profiles.

Program developers call this is an "assess, group, teach" approach to intervention. Specifically, the process of BURST supplemental instruction proceeds as follows:

Assess: At the beginning (and again at the middle) of the school year, all students at relevant grades in a school are administered a mobile version of the DIBELS Next® assessment (described at: https://acadiencelearning.org/ DIBELS_Next_Info.pdf). *DIBELS Next®* is a criterion-referenced assessment that has grade-level, time of year test forms that measure student achievement in the areas of letter naming fluency, phoneme segmentation fluency, nonsense word fluency for letter sounds and whole words, oral reading fluency, and passage reading comprehension. The test developers have established cutoff scores on each of these forms that empirical evidence shows are predictive of future performance on the assessment. Using these cutoffs, students are classified as falling well below benchmark, below benchmark, at benchmark, and above benchmark.

Group: Immediately after universal screening is completed, students' item responses from the *DIBELS Next®* assessment are transmitted via the internet to Amplify Inc. for scoring. Amplify scores the assessments,

Figure 1: BURST Curriculum Strands



which identify students performing well below, below, at, or above DIBELS grade-level/time-of-year benchmarks for particular early literacy skills. Schools then access student scores and choose the students they wish to receive BURST treatment.

With students identified for treatment, Amplify then uses a proprietary software algorithm to group identified students into subgroups with similar assessment profiles. These subgroups are then assigned to specific 10-day cycles of instruction within the BURST curriculum, where instructional cycles are matched to students' assessment results. Figure 1 (above) provides an overview of the BURST curriculum that is assigned to students as a result of the assessment/grouping process. Each hexagon in the figure is a strand of the BURST curriculum for which multiple 10-day instructional cycles have been developed. Strands on the left are meant to be taught before those on the right. In this design, the BURST algorithm seeks (under constraints) to assign nominated students to a small group offering instruction on the lowest level in the skill sequence that students in the group have yet to master (as judged by to *DIBELS Next®* benchmark proficiency scores).

Teach: Once student groups are assigned to a given BURST strand, teachers are given program-developed lesson materials to provide students with targeted instruction. Instruction proceeds in 10-day lesson cycles. Each daily lesson in a ten-day cycle is built around a template designed for use in a 30-minute intervention session. Each lesson template is a rough "script" for a day of instruction (along with associated instructional materials).[1] Lesson templates are organized so that a teacher will introduce a skill, model how to apply that skill, and then give students time to practice skill application. Near the end of any given 10-day BURST cycle, students in a functioning BURST group are assessed using a curriculum embedded progress-monitoring assessment. Results from these assessments are also sent to Amplify, Inc. for processing and are used to assign students to subsequent 10-day instructional cycles based on assessment results. In schools where the program is a regular feature of supplemental instruction, program developers assume that any student assigned to

[1] A complete list of instructional materials used in the BURST program can be found at https://burstbase.net/.

BURST treatment will experience six 10-day BURST cycles between universal screenings. So, in a roughly 90 day interval between universal screenings, Amplify suggests that identified students receive roughly 60 days of BURST instruction, and if each lessons takes 30 minutes, that becomes roughly 30 hours of supplementary instruction per semester.

Logic of Program Implementation

Program implementation requires use of a uniform set of assessment and grouping services provided by Amplify along with uniform curriculum and instructional materials developed by the company. Importantly, schools implementing BURST have discretion over the amount and kind of training and support services they receive over time from Amplify. In addition, Amplify assumes the BURST program will be implemented in schools that differ in school-level staffing configurations, schedules, student composition, and resource constraints. Therefore, Amplify does not tightly specify who will manage the BURST program at a given school, who will teach BURST lessons to identified students, how many students will be allocated to BURST treatment, or how many BURST groups are formed. The logic of program implementation is thus a mix of highly specified elements (like assessments, curriculum, and lesson plans) coupled with flexibility in implementation support and program delivery to students. These elements are now described.

Purchase of Training and Support for DIBELS Next® Implementation: Schools that have not subscribed to Amplify's *mCLASS: DIBELS Next®* services prior to subscribing to BURST will ordinarily purchase this service and one or more training options to learn about the *DIBELS Next®* assessment and how it can be used to improve early grades reading instruction (the menu of *mCLASS: DIBELS Next®* training choices can be viewed at https://www.amplify.com/assessment/mclass-training).

Purchase of BURST®: Reading Services and Materials: In addition, schools that purchase BURST (typically) pay an annual per-student license fee for digital access to the BURST program's customized curriculum modules and grouping and reporting services, which are delivered via secure internet connection. Teachers delivering BURST instruction to students also will use a variety of materials, including various flash cards, portable white boards and markers, counting chips, stickers, a puppet, and more. Teachers can download many of these materials from the BURST internet site, or a school can buy each teacher a kit with these materials.

Purchase of Training and Supports for BURST®: Reading Implementation: Schools that purchase BURST also purchase a training and implementation support package from Amplify. The base training program includes a one-day, on-site session hosted by Amplify's professional services staff. This training is designed for teachers or interventionists, and training can be delivered to all school personnel who will implement the BURST program or just a cadre of teachers who will, in turn, train other teachers at a school site. Regardless of the attendees, the initial training session focuses on how to implement the BURST program with fidelity, covering such topics as how to administer formative assessments, access sequences of lessons through the web-based interface, deliver instruction, and monitor success based on the program's curriculum-embedded formative assessments. Schools subscribing to BURST can also contract for additional, on-site consultations with Amplify's educational support team. A typical consultation is a day in length. During this day, a member of Amplify's educational support team will visit a school to offer formative guidance and motivational support to users. The site visit will include meetings with the principal and any other personnel in charge of program implementation at a school, direct observations of BURST small group instruction in which Amplify staff use a fidelity checklist to rate the quality of BURST instruction observed, and a focus group meeting with all BURST instructors designed to troubleshoot implementation issues and provide formative feedback to school personnel.

Organization and Management of BURST Program at Schools: Crucially, Amplify expects (and allows) schools to use a variety of organizational and personnel arrangements to manage program implementation and deliver instruction to students.[2] These organizational arrangements are now described.

- ***Program management***: Amplify recommends that schools using BURST appoint a program coordinator to manage the program. Operating the program involves scheduling DIBELS assessments, managing the data transactions between schools and Amplify that organize instructional delivery, and coordinating delivery of BURST instruction within a school. The program manager can be the school's principal, a special programs coordinator, or a member of the faculty, and once the BURST program is up and running at school, Amplify predicts that the coordinator will spend 2-5 hours per week managing the program. Amplify does not exercise direct control over schools' choice of program manager or the activities that manager engages in.

- ***Teaching Personnel:*** Amplify expects that schools will use BURST as a supplementary instructional program and that BURST lessons will be delivered to small groups of students in a pullout or push-in setting. Within schools, BURST lessons can be taught by a trained "interventionist" (usually a reading specialist), by classroom teachers with elementary teaching certificates, or by paraprofessionals. Amplify does not exercise direct control over who teaches BURST lessons, the setting in which lessons are taught, the scheduling of BURST lessons, or the frequency with which BURST lessons are taught to students.

- ***Instructional Grouping Arrangements:*** Amplify allows school personnel to influence the software routines used to place students into BURST groups. In all schools, schools decide how many (and which specific) students to serve with BURST and they also specify a preferred BURST group size (based on resource constraints). Schools then input into the BURST grouping function the specific students to be served, the number of groups to be formed, and a preferred group size. Amplify recommends that all students scoring below or well below DIBELS grade/time of year benchmarks be included in the group formation process and that the preferred group size be set at 4.5 students per group. However, Amplify also recognizes that schools vary in how many students are below these benchmarks and their capacity to serve different numbers of students. Thus, across any set of BURST schools, the number of students identified for treatment and the number of BURST groups operating will vary.

- ***Lesson Assignments***: Amplify recommends that BURST groups be formed across classrooms (and even grade levels). This is because the proprietary grouping algorithm used in BURST will be better able to form homogeneous groups of students with similar assessment profiles if larger numbers of groups are being formed. However, Amplify exercises no direct control over school decisions about this matter, and in many settings, groups will be somewhat heterogeneous. Under these conditions, if the group has three students, BURST will target instruction based on the student with earliest skills need; if there are 4-5 students in a group, Burst targets instruction based on the student with second earliest skill needs; if there are 6 students in a group, Burst will target instruction based on the student with third earliest skill needs.

Instructional Delivery: Once groups are formed, these groups are expected to progress through ten-day BURST instructional cycles. As discussed earlier, each daily lesson is built around a template designed for use in a 30-minute intervention session. Each template is a rough "script" organized so that a teacher will introduce a skill, model how to apply that skill, and then give students time to practice skill application. Near the

[2] Although Amplify expects (and allows) diverse arrangements for program implementation, it does provide guidance on these matters (see https://burstbase.net/faqs/).

end of any given 10-day BURST cycle, students in a functioning BURST group are assessed using a curriculum embedded progress-monitoring assessment. Results from these assessments are also sent to Amplify, Inc. for processing and are used to assign students to subsequent 10-day instructional cycles based on assessment results. This repeated teaching of BURST cycles continues until the next universal screening.

Summary

To summarize, *BURST®: Reading* is an intervention program designed to provide supplemental instruction in early literacy skills to elementary grades students who have not yet mastered grade-level reading skills. The program uses universal screening to identify students performing below grade level/time of year benchmarks on the *DIBELS Next®* assessment and then uses results from this screening to place students into small groups for targeted supplemental instruction. Amplify provides a number of resources that can be used in teaching BURST lessons, including lesson "scripts" and materials kits, and it provides a menu of training and support options that schools can use to implement the program faithfully. However, Amplify also recognizes that the BURST program will be implemented in schools with differing staffing configurations, schedules, student composition, and resource constraints. As a result, Amplify provides schools considerable flexibility in who manages the BURST program at a given school site, who teaches BURST lessons to identified students, how many students are allocated to BURST groups, and how many BURST groups are formed.

3 | APPROACH AND RESEARCH QUESTIONS

This section briefly describes the design of the *BURST®: Reading* efficacy trial and the research questions addressed by the study given that design. The efficacy trial used a cluster-randomized field design in which intact schools were randomly assigned to treatment and control conditions. Control schools received free access to DIBELS testing services for students in grade K-3 while treatment schools received free access to DIBELS and BURST services in grades K-3. The study followed treatment and control schools over a four-year period, during which time Amplify collected data on program operations in treatment schools and on the early literacy achievement of students in grades K-3 in both treatment and control schools.

Given the study design, researchers at the University of Michigan explored research questions in three domains. An initial set of questions were about the samples of schools recruited and retained in the sample and the extent to which random assignment of schools to treatment worked as expected. Here, the questions are about the characteristics of schools in the study, baseline equivalence of treatment and control schools on pre-treatment covariates, and patterns of attrition from the study. A second set of questions concern how schools assigned to receive the BURST treatment actually implemented that program, including questions about the numbers and kinds of students who actually received BURST instruction in treatment schools and the frequency of that instruction. A third set of questions concern program effects on student achievement, including questions about BURST program effects on early literacy achievement averaged across all students, across students that the program's theory of action suggests should be prioritized for BURST instruction, across students who were exposed to the BURST program for varying lengths of time during the study, and across students who were in schools that could be expected on the basis of pretreatment covariates to comply with their assigned treatment group status. These questions are discussed in more detail below.

THE RECRUITED SAMPLE AND SUCCESS OF RANDOM ASSIGNMENT

Since Amplify views *BURST®: Reading* as a *school-level* intervention program, this study was designed as a cluster-randomized field trial in which intact schools were randomly assigned to treatment or control conditions. Recruitment of schools to the study began in October of academic year (AY) 2012-2013 and concluded in November of AY 2013-2014 (just after the September 2013 launch of the study). Random assignment occurred in six steps during the recruitment stage. Once a group of schools was recruited into the study at a given time point, schools in that group were randomly assigned to treatment or control status using the following procedure. First, recruited schools were grouped by state. Next, matched pairs of schools were created within states. Then, schools within pairs were randomly assigned to the treatment or control condition. Importantly, there was considerable attrition during the recruitment stage of the study, but *not* after the study was launched. In particular, 92 schools in total were recruited into the study, but 40 left (or were dropped from) the study prior to its launch, leaving a retained sample of 52 schools.

The process of recruitment (and attrition) will be discussed in more detail in Chapter 4 of this report, where we also discuss several research questions related to sample recruitment, retention, and random assignment. One question discussed in Chapter 4 concerns the "external validity" of the current field trial. As Kern et al. (2016) note, education researchers are increasingly interested in the extent to which the samples of participants in randomized field trails represent some target population of interest, that is a population to which the results of the study are meant to be generalized.[3] In this study, an attempt was made to recruit schools with more than 50% of students eligible for participation in the federal free or reduced price lunch (FRL) program.

[3] Kern, H. L., Stuart, E. A., Hill, J., & Green, D. P. (2016). Assessing methods for generalizing experimental impact estimates to target populations. *Journal of research on educational effectiveness*, *9*(1), 103-127.

We are therefore interested in the extent to which this study goal was achieved. In addition, we are interested in another target population to which we might want to generalize our study findings, namely, the kinds of schools that normally purchase *BURST®: Reading*. In what follows, we call this the BURST user group. These interests lead to the following research question:

> **RQ1 (Characteristics of the Sample)**: What kinds of schools (and students) participated in the current study? How did the sample of recruited (and retained) schools compare to the population of *all* U.S. schools serving grades K-3? Did the study meet its goal of recruiting schools with greater than 50% participation in the free/reduced price lunch program? Also, how did the sample of schools assigned to treatment compare to the BURST user group of schools that purchased and were implementing the *BURST®: Reading* program as of AY 2013-2014?

Beyond questions about sample composition, we are also interested in potential threats to the "internal validity" of the current study that might have arisen during the recruitment and random assignment phases of the study. Threats to internal validity could be a result of: (a) the chance that random assignment procedures used in this study failed to completely "control away" differences in baseline characteristics of treatment and control schools (and students); and (b) possible differential attrition after random assignment that could also lead treatment and control schools remaining in the study to differ on characteristics related to the study's outcome. These considerations lead to two additional research questions:

> **RQ2 (Differential Attrition)**: To what extent did rates of attrition from the study differ across schools assigned to treatment and control conditions?

> **RQ3 (Baseline Equivalence)**: To what extent did the random assignment of schools to treatment and control conditions result in schools (and students) in the two conditions being similar on baseline covariates that are potentially related to students' early literacy achievement? In particular, given attrition from the study, to what extent were schools (and students) retained in treatment and control conditions similar on baseline covariates potentially related to students' early literacy achievement?

IMPLEMENTATION OF BURST®: READING IN TREATMENT SCHOOLS

A second domain of research questions concern how the BURST program was implemented in schools assigned to treatment. Questions about implementation are important for several reasons. To begin, the Institute of Education Sciences (IES) funded the current study as "an initial *efficacy* study of a widely used intervention" (emphasis added). However, the current study was not an efficacy trial in the conventional sense of that term. By the usual definition, an efficacy trial is mounted under well-controlled circumstances, where samples are carefully chosen and program implementation is closely controlled to assure fidelity of implementation. By contrast, IES has a more liberal definition of an efficacy trial. To begin, it requires efficacy trials to be conducted in "authentic education settings" (not well-controlled laboratory or clinical settings). Moreover, IES allows efficacy studies to be conducted under a mix of "ideal" and "routine" circumstances (as opposed to the ideal conditions usually imposed in conventional efficacy trials). That mix of conditions occurred in the present study. Schools in the treatment group were given supports that are not ordinarily available to users of DIBELS and BURST—namely *free* services and training designed to assist schools in use of the DIBELS and BURST—making this an ideal condition for schools. However, Amplify did not force schools to participate in these trainings. In addition, Amplify offered schools additional supports during the implementation phase of the study (in the form of free visits to school sites by Amplify educational support staff), but once again, schools were free to take advantage of this service or not. Finally, Amplify gave treatment schools wide discretion over instructional delivery, in particular, how many students in a school would receive

BURST instruction, the size of BURST groups used to deliver that instruction, and the number of 10-day cycles of BURST instruction offered to students. In this sense, then, the current efficacy trial was conducted under a mix of ideal and routine conditions. Schools were provided free training and support, but could choose how much to take advantage of these supports, and schools had considerable latitude (as they do under "routine" conditions of implementation) to offer instruction in ways that fit their circumstances.

All of this suggests the need to carefully examine how the BURST program was implemented in schools assigned to treatment in this study. An initial question concerns the training and implementation support services delivered by Amplify to treatment schools:

> **RQ4 (Delivery of Training and Implementation Services to Treatment Schools):** To what extent did treatment schools take advantage of the free DIBELS and BURST training services offered by Amplify? Also, to what extent did treatment schools take advantage of the follow-up support services offered by Amplify?

A second set of questions concerns the "fidelity" of implementation of BURST instruction within treatment schools. An initial set of questions concerns the percentage of students who actually received BURST instruction, as well as the average number of BURST cycles a student received. The questions here are:

> **RQ5 (Provision of Any BURST Instruction to Students):** To what extent did schools in the BURST treatment group provide BURST instruction to students? That is, what percentage of students in treatment schools were placed into BURST groups, and how many cycles of BURST instruction did students receive?

Because this is an efficacy trial, we were further interested in how provision of instructional services in treatment schools compared to "ideal" and "routine" implementations of the program. Under "ideal" conditions, Amplify recommends that an identified student receive six BURST cycles per semester (or 12 cycles per year). Thus, an additional research question asked was:

> **RQ6 (Provision of Ideal Doses of BURST Instruction to Students):** To what extent did schools in the BURST treatment group meet the standard of "ideal" provision of BURST instruction to students? That is, what percentage of students in treatment schools received 6 cycles per semester and 12 cycles per year of BURST instruction?

An additional question concerns whether provision of BURST instruction in treatment schools differed from what one might observe under "routine" conditions of implementation. To address this question, we acquired additional data from Amplify's *mClass* data system, a data systems that routinely collects data relevant to our questions for *all* schools that purchase the BURST program. In the current study, we compared relevant implementation data collected on schools in the treatment group to relevant data on program implementation at 671 schools that had purchased and were using BURST as of AY 2016-2017. Using these data and methods discussed in more detail in Chapter 5, we addressed the following question:

> **RQ7 (Provision of Routine Doses of BURST Instruction to Students):** Did schools in the BURST treatment group provide BURST instruction to proportionally more (or fewer) students than schools that purchased BURST and were using it under routine operating conditions? In addition, were schools in the BURST treatment group more (or less) likely to provide the ideal number of cycles of BURST instruction to students as compared to schools that purchased BURST and were using it under routine operating conditions?

Finally, we were interested in the extent to which instructors of BURST groups followed the lesson "scripts" that are a central feature of the BURST program. To examine this issue, members of Amplify's education support group were trained to use an implementation fidelity checklist to rate the extent to which various features of a BURST lesson script were used by the instructor, and these personnel used these checklists when they observed BURST groups during their site visits to a selected group of schools. This leads to another question:

> **RQ8 (Use of BURST Lesson Routines in BURST Instruction):** To what extent did the instructors of BURST groups follow the prescribed elements of the BURST lesson template?

PROGRAM EFFECTS ON STUDENT ACHIEVEMENT

A final domain of research questions asks about the effects on student achievement attributable to a student's enrollment in a school assigned to BURST treatment, where student achievement is measured in this study by the Renaissance *Star Early Literacy®* assessment.[4] Our research questions about student achievement reflect the fact that schools (not students) are the unit of treatment. As a result, in the current study, "treatment" was measure as a dichotomous variable (where a school either is or is not assigned to treatment). The reader will note that, in this approach, the "treatment" condition in the study is the adoption (not implementation) of BURST, and in this approach, any student in attendance at a BURST school is seen as having been exposed to treatment. In what follows, we ask a series of questions about who benefitted from exposure to treatment in the RCT and about whether any benefits observed are attributable to a school's assignment to BURST. Our statistical approach to estimating these attributable effects is described in detail later in this report and in Appendix A. In this chapter, the task is simply to state the research questions in everyday language, not the formal language of statistical hypothesis testing.

The first question we ask about student outcomes examines the effect on students' achievement of being in a school assigned to the BURST treatment group when that effect is averaged across all students. The question is relevant to the current study because BURST is designed as a school-level instructional intervention, and it is reasonable to ask what the benefit of being in a BURST school is for students in general. In the research literature, this effect (of matriculating in a BURST school) is called the "average treatment effect" or the "intent to treat" effect. Importantly, it is *not* the effect on achievement expected for students who received any BURST instruction in a school, nor is it the expected effect on achievement of having received an "ideal" dose of BURST instruction. Instead, it is simply the estimated effect on students' achievement after averaging across all students who were in the BURST treatment group. Stated informally, the "average treatment effect" question is:

> **RQ9 (Effect of BURST Averaged Across all Students):** What was the effect on students' early literacy achievement for *all* students enrolled in a BURST school?

The "average treatment effect" that is the focus of RQ9 is useful to researchers and policy makers facing accountability or other reporting requirements who might want to know the extent to which the average achievement in a school will increase if that school adopts BURST. But we are also interested in whether being in a BURST school has more or less benefits for different subgroups of pupils. In particular, BURST is a supplemental reading program designed for struggling readers, and the program is designed to identify struggling readers (using DIBELS testing) and then to provide such students with indicated instruction. Because of this, we are particularly interested in the extent to which struggling readers attending BURST schools experience boosts to their early literacy achievement. In Chapter 6, we present two estimates of this ("conditional average treatment effect"), one for students who entered a BURST school having scored below the relevant DIBELS grade/time-of-year benchmark at time of entry, and another for students who at any point in

[4] The research questions listed in this section are the confirmatory research questions about student outcomes listed with Society for Research on Educational Effectiveness *Registry of Efficacy and Effectiveness Studies* (www.sreereg.org).

the study period scored below DIBELS benchmarks at a given grade/time of year. Informally, this combined research question can be stated as:

> **RQ10a (Effect of BURST on the Achievement of Struggling Readers):** What was the average effect of being enrolled in a BURST school on the early literacy achievement of students whose observed DIBELS composite score at time of entry into a treatment school was below or well-below the DIBELS grade-level/time of year benchmark at that time point?

> **RQ10b (Effect of BURST on the Achievement of Struggling Readers):** What was the average effect of being in BURST school for students who had a composite DIBELS score below or well-below a grade/time of year DIBELS benchmark at any time in the study?

We are also interested in possible boosts (or decrements) to achievement attributable to be being in a BURST school even if a student is *not* a struggling reader. This group of students is defined in the current study as students who performed at or above DIBELS benchmarks at entry into the study and students who continued to score at or above benchmarks at all succeeding time points at which they were observed. One can think of various reasons why these students might benefit from being in a BURST school even if they do not receive any BURST instruction. Such students might, for example, benefit from being in a BURST school because struggling readers in their school become better learning partners during regular reading instruction (a so-called "peer" effect) or because a teacher in a BURST school can more easily accelerate the pace of regular instruction when struggling readers' receive supplemental instruction. Alternatively, mounting BURST instruction could negatively affect higher achieving students' outcomes, especially if mounting BURST instruction somehow drained resources from the regular reading program. Stated informally, then, our research question is:

> **RQ11 (Effect of BURST on the Achievement of Readers at or above Benchmark):** What was the average effect of being in BURST school for students who at the start of the study or at all points in the study had a composite DIBELS score at or above grade/time of year benchmarks?

An additional research question asks about the effects on students' early literacy achievement of continuous exposure to treatment. Questions about length of exposure to treatment are possible in the current study because the BURST efficacy trial was conducted over a four-year period. In Section 6 of this report, we report descriptive data on treatment effects experienced by students after one, two, three, and four years of study participation. We also take advantage of the longitudinal data collected for this study to ask a more specific question about the effects of continuous exposure to treatment on students' early literacy achievement, especially for struggling readers. The idea here is that struggling readers who matriculate continuously at a BURST school will experience cumulative effects of exposure to treatment—both through increased chances of being allocated to BURST groups for supplemental instruction and/or by being exposed over a longer period of time to potential "peer effects." In this report, we define continuous exposure to treatment as continuous enrollment at a BURST school over a three or four year period, and we define struggling readers as students who ever scored below DIBELS benchmark during their time in the study. With these definitions in hand, our research question can be stated informally as:

> **RQ12 (Effect of Continuous Exposure to BURST Treatment on Struggling Readers Achievement):** What was the average effect on the early literacy achievement of struggling readers who were in continuous attendance at a BURST school for three to four years?

Our final research question is about the effect on all students' achievement of attending a school that has at least some chance of offering BURST instruction to students. At various points in this report we have noted that Amplify does not require BURST purchasers to actually offer BURST instruction to students, and in the current study, that was also the case. Therefore, the possibility exists that at least some schools assigned to treatment in the RCT will not comply with their treatment assignment—that is, will not offer BURST instruction to students. We could, of course, use data from the study to identify any non-compliant schools, but schools not offering BURST instruction might have done so because they found the program had no effects on their students. On this view, the observed provision of BURST instruction to students is endogenous to treatment assignment. To work around this issue, we used an "out of sample" statistical model to predict non-compliance in treatment schools based on their pre-treatment covariates. The prediction model was a machine-learning Bayesian adaptive tree regression model trained on the *mClass* data for the AY 2016-2017 BURST user group—that is, the population of all schools serving students in any grade K through 3 that had previously purchased and were using BURST (but were not in the RCT). This prediction model is described in more detail in Appendix B. The model had as predictors of compliance a number of pre-treatment covariates for schools (including prior levels of student achievement at schools, a number of school-level student composition and structural variables, and various district-level funding and community location variables). These variables were used to predict the percentage of students in a school who received 12 or more cycles of BURST instruction in AY2016-2017. The number of cycles was set at 12 because this represents the "ideal" dose of BURST instruction for students identified as in need of BURST instruction. We then applied the regression coefficients from this model to AY 2012-2013 values on the covariates for schools in the BURST study, giving us for each school in the study a "compliance" prognosis scaled as the predicted percent of students expected to receive "full" BURST treatment. We then defined any school with a predicted percentage of less than 1% to be an expected "never complier" and any schools with a predicted percentage of greater than 1% to be a "complier." Using these data, we addressed the following research question:

> **RQ13 (Effect of Being in a School with an Expectation of Compliance to Treatment Assignment):**
> What was the average effect on early literacy achievement for students enrolled in a BURST school predicted to comply with treatment assignment?

SUMMARY

We have just discussed research questions about the average effects on students' early literacy achievement attributable to their enrollment at schools assigned to BURST treatment. As discussed, we are interested in these effects averaged over all students and averaged across subgroups of students, including struggling (and non-struggling) readers, students who were continuously enrolled at BURST schools for three or four years of the study, and for students who were enrolled in schools that were predicted (on the basis of an out-of-sample prediction model) to comply with their BURST treatment assignment. We also discussed a set of related research questions that will: (a) allow the reader to assess threats to internal validity of the RCT that might have arisen due to attrition of schools from the study; (b) provide the reader with information relevant to the external validity of the study; and (c) allow the reader to understand the extent to which actual provision of BURST instruction to students enrolled at treatment schools resembled what would be expected under "ideal" or "routine" conditions of implementation.

4 | Sample

This chapter describes several issues related to the sample of schools that participated in the BURST efficacy trial. In a first section, we describe the steps taken to recruit schools into the study and the steps to taken to assign recruited schools to treatment and control conditions. In a following section, we describe the baseline equivalence of schools initially assigned to treatment and control during the recruitment phases of the study, the attrition of schools that took place at this phase of the study, and whether attrition affected baseline equivalence of schools in the remaining analytic sample. In a final section, we compare the recruited and retained samples to two additional populations of schools—the national population of schools serving students in any of grades K-3 and the population of schools serving any of grades K-3 that were subscribed to BURST in AY 2016-2017.

RECRUITMENT AND RANDOM ASSIGNMENT

Recruitment of schools began in August of the AY 2012-2013 school year and concluded in November of AY 2013-2014. Because the study was designed to assess the effects of BURST adoption on students' *early* reading achievement, only schools containing grades K-3 were allowed in the study, and a goal was set to recruit schools from this pool that also had greater than 50% students eligible for the federal government's free and reduced price lunch program. The initial plan was to recruit approximately 50 such schools into the study so that data collection could be launched in AY 2012-2013. However, delays in recruiting postponed the launch of the study to Fall of AY 2013-204.

Scheduling of Recruitment and Random Assignment: Table 4.1 (next page) provides a summary of the pace of recruitment activities during this time period. Notice from the table that schools were recruited into the study in batches. Over the course of recruitment, 92 schools located in 11 different states agreed to participate in the study, but of these 92, only 52 schools (in 9 states) ended up participating in the study at launch. In what follows, we call the 92 schools the "recruited" sample, and the remaining 52 schools the "analytic" sample.

More than half the attrition of schools from the study came from schools recruited in the October 2012 batch of recruited schools. Of this group, 24 of 26 schools either withdrew from the study or were dropped by study researchers under a policy described in more detail below. Another 66 schools were recruited into the study between January 2013 and November 2013, and of these, 16 schools either withdrew from the study prior to launch or were dropped from the study by researchers under a policy described in more detail below. A good share of the attrition in the study was due to the delayed start, but attrition also occurred when schools received notification of their status as control schools in the study. However, all of this attrition occurred *prior to* the launch of the study. Thus, the 40 schools that withdrew or were dropped from the study never received services or participated in data collection as part of the study. Only the 52 schools in the analytic sample were provided services and participated in data collection.

Table 4 also shows that treatment assignments were made at six distinct allocation times. At each allocation point, schools were grouped into pairs or triples prior to random assignment using a procedure (described in the next section) that created allocation groups with similar demographic characteristics and student achievement histories. To enable this procedure to make use of state testing results from years prior to the study, groupings were made within state whenever possible. Indeed, only 2 of 45 pre-randomization pairs or triples crossed state lines. All told there were 43 pairs and 2 triples, the latter occurring at allocations *ii* and *iv*, when the total number of schools to be assigned was odd. One school in a triple (in allocation batch *ii*) was assigned to intervention, while two were assigned to treatment in another triple (in allocation batch *iv*). Thus, Table 4.1 | Recruitment and Random Assignment of Schools

Recruitment Batch	Date of Random Assignment	Number of Districts in Batch	Number of Schools in Batch	Number of Pairs for Random Assignment	Number of Triples for Random Assignment
i.	Oct. 2012	15	26	13	–
ii.	Jan. 2013	1	13	5	1
iii.	May 2013	5	26	13	–
iv.	1 July 2013	4	13	5	1
v.	15 July 2013	2	4	2	–
vi.	Nov. 2013	1	10	5	–
Total	–	27	92	43	2

on the whole, probabilities of assignment to the intervention condition ranged from 1/3 (for the batch *ii* triple), to 1/2 (for any school in a pair), and 2/3 (for the batch *iv* triple).

Formation of Pairs and Triples Used in Random Assignment: The pairs and triples formed prior to random assignment were constructed using the optimal matching procedure of Greevy et al. (2012).[5] The procedure is designed to minimize discrepancies across paired (or tripled) schools on a multivariate distance measure that included various pre-treatment characteristics of schools, including data on from state records on: school size, student demographics, and state test results of recent 3rd grade cohorts. Because the optimal matching mechanism could make use of only a limited number of these variables, the initial allocation to condition in October 2012 tested a number of combinations and weightings of the variables in a dry run of the match-and-then-randomize procedure used for later batches. Results from these dry runs were assessed in terms of the "on-average" similarity (on matching variables) of pairs of schools. This assessment found it favorable to match on the average size of grade-level cohorts in grades K-3, percent of students identified as White; free or reduced price lunch eligibility; proficiency rates in reading and writing, separately averaged over 2010, 2011 and 2012 third grade cohorts; and, for each of the preceding variables that were incompletely recorded, a {0,1}-valued indicator of data availability. The dry run assessment also supported giving the cohort size and averaged proficiency rate variables twice the weight of the demographic and missingness indicators. No additional dry run assessment was made for allocations 2-6, although the procedure was adapted by exclusion of the writing proficiency variable, which was frequently unavailable; by replacing the grade cohort size variable with an average of total enrollments in 2010, 2011 and 2012; and by an inadvertent substitution of proportions of third graders testing as Proficient for proportions testing as Proficient or Advanced. Within each recruitment batch, once optimally matched pairs and triples were identified, schools in each pair or triple were randomly assigned to treatment or control.

RESULTS OF THE RANDOM ASSIGNMENT PROCESS

In this section, we discuss the results of this random assignment process in three steps. We begin by showing that samples of treatment and control subjects were well-balanced on important pretreatment covariates immediately after random assignment. We then show that despite high levels of attrition of schools from the

[5] Greevy Jr, R. A., Grijalva, C. G., Roumie, C. L., Beck, C., Hung, A. M., Murff, H. J., ... & Griffin, M. R. (2012). Reweighted Mahalanobis distance matching for cluster-randomized trials with missing data. *Pharmacoepidemiology and drug safety*, *21*, 148-154.

Table 4.2 | Data on Baseline Equivalence of Treatment and Control Schools for Recruited Sample (n=92) and Analytic Sample (n=52):

	Recruited Schools			Analytic Sample		
AY 2011-2012 Variables Used in Matching	Control	Treatment	Difference	Control	Treatment	Difference
Average No. of Pupils in Each of Grades K-3	81	68	-13	74	61	-13
Percent White	36	35	-1.5	58	57	-0.65
Pct. FR Lunch	77	74	-3.2	75	74	-0.49
G3 Reading Proficiency	60	58	-1.6	57	54	-3.2
AY 2011-2012 Variables Not Used in Matching	Control	Treatment	Difference	Control	Treatment	Difference
Pupils Per Teacher	22	23	1.9	21.8	20.6	-1.2
Pct. Special Education	11.3	11.7	0.4	11.0	12.4	1.3
G3 Math Proficiency	53	52	-0.95	57	54	-3.1
AY 2012-2013 Variables (Not Used in Matching)	Control	Treatment	Difference	Control	Treatment	Difference
G3 Reading Proficiency	57	60	3.9	54	57	2.2
G3 Math Proficiency	60	62	2	60	58	-2.3
Urban	46	45	-0.37	33	32	-0.15
Suburban	21	18	-2.4	37	34	-2.9
Rural or Town	33	36	2.8	31	34	3.1

study, attrition rates were roughly similar in the treatment and control conditions. Finally, we show that despite high attrition, the analytic sample was balanced on pre-treatment covariates in the same way it would be expected had attrition not occurred.

Much of the data for this discussion is presented in Table 4.2 (above). The first column of the table lists the variables on which the schools are being compared while subsequent columns present control and treatment group means, and differences of these means. Columns 2-4 compare all 92 schools randomly assigned as part of the study, while columns 5-7 make similar comparisons restricted to the 52 schools constituting the analysis sample, i.e. those schools that did not leave and were not dropped from the study. The table presents selections of the AY2011-2012 variables used in matching, other AY2011-2012 school variables that were not used in matching, and AY2012-2013 variables that became available only after matching and random assignment were completed. Schools in the analytic sample began receiving services in AY 2013-2014. Means presented Table 4.2 (columns 2, 3, 5 and 6) are weighted for school size. In addition, for comparability between the two columns of figures, weights used to average over the control group (columns 2 and 5) incorporate a factor equal to the odds of assignment to the treatment group. Thus control schools randomized within pairs are simply weighted by school size, as their counterparts in the treatment group are; but in the randomization block of three schools that assigned one to control and two to treatment, the single control school receives a weight equal to twice its size, so that treatment and control schools belonging to this block contribute similar shares of the total "mass" represented by the weighted means over the treatment and control groups. Another randomization block assigned two schools to control and one to treatment, and each of these control schools carried a weight equal to half its size. Columns 5-7 use the same weighting scheme to compare the 25 control and 27 treatment schools that constitute the analytic sample for this study.

Baseline Equivalence of Treatment and Control Groups in Recruited Sample: If we consider all 92 schools ever recruited to the study and randomized, Table 4.2 supports the inference that treatment and control groups showed "baseline equivalence" across many pre-treatment covariates. Consider, for example, the pre-treatment covariates measured at AY 2011-2012 and used in the optimal matching process. As the first panel of Table 4.2 shows, there were only small or moderate differences across treatment and control groups in the 92 school sample, and unsurprisingly, these differences were *not* statistically different from 0 ($\chi^2 = 8.1$ on 7 d.f., p=0.3). Differences between treatment and control groups in the recruited school sample were also small on a set of variables *not* used in the matching procedure but available for the same time period were are shown in the second panel. Here, net of differences on matching variables, the combined differences statistic (Hansen & Bowers, 2008) was $\chi^2 = 1.7$ on 7 d.f. (p=1.0). Yet a third collection of pre-treatment variables, which became available only after random assignment, measured school characteristics in AY 2012-13 (the year prior to study launch). These variables are shown in the third panel of Table 4.2. Here too there were only small differences across treatment and control groups in the recruited sample, those difference being on par with what the cluster randomization procedure would be expected to produce ($\chi^2 = 6.2$, net of differences on earlier baseline variables, with 12 d.f.; p=0.9). Again, these are baseline comparisons of *all* schools assigned to treatment or control, without regard to whether a school stayed in the study or was among those that left the study, and the hypothesis under test posited no differences across treatment and control groups other than those that would be expected under random assignment.

Attrition of Schools from the Recruited Sample: Although there was strong evidence of baseline equivalence of schools after random assignment, we have seen that there was substantial "attrition" of schools from the study. If one takes schools as the unit of analysis, there is an attrition rate of about 43.5%. If one takes students as the unit of analysis (i.e., weights attrition by school size), the overall attrition rate was 47.4%. Attrition mostly occurred among schools assigned to treatment during the first two of 6 recruitment cycles, and all schools that dropped out of the study did so within a few weeks of receiving their random assignment but well prior to the delivery of services and collection of study data. The most common reason schools left the study was receipt of control assignment, but the research team followed a recruitment/retention policy that limited differential attrition resulting from this process. In particular, in recruiting that took place after October 2013, the research team informed districts that if any school recruited to the study were to cease participation in the study, the study team would discontinue delivery of services to other schools in a single-district pair or triple to which that school belonged. All but one of the schools recruited during this latter period came from a district contributing other schools, and because the pairing procedure matched schools within districts when possible, districts were almost certain to have schools in both study arms. As it happens, the one school that was the sole representative of its district was assigned to control (as part of batch *iii*) and dropped out of the study shortly afterward. Its counterpart was retained, but each of the other 39 schools that are counted in our attrition group were removed alongside their matched counterpart. That includes 24 of the 26 batch *i* recruits, all 13 of the batch *ii* recruits and 2 of the 13 batch *iv* recruits. As a consequence of this recruitment/retention policy, treatment schools' attrition rate was, in the end, only slightly lower than that of control schools. Indeed, if one weights for school size and, in the case of the control group calculation, the reciprocal of the *ex-ante* odds of assignment to that group, treatment schools' attrition rate was only slightly lower than that of schools control schools. By this calculation, attrition for control schools was 46% as opposed to 48% for treatment schools. This 1.5% difference in attrition rates was *not* statistically significant and qualifies the current study for review by the *What Works Clearinghouse.*[6]

[6] In the weighting procedure just described, schools randomized within *pairs* are simply weighted by size; the one school randomized to control as part of a triple with two treatment schools is weighted by twice its size; and the two control schools randomized to control as part of a triple along with one treatment school were weighted by one half of their sizes. Note that weighting for reciprocal odds of assignment but not school size leads to a difference of attrition rates of

Baseline Equivalence of Treatment and Control Groups in Analytic Sample: The attrition process just described left us with an analytic sample of 52 schools, 27 of which had been randomly assigned to BURST treatment, and 25 of which had been randomly assigned to the treatment condition. Beginning in Fall of AY 2013-2014 these schools began receiving services appropriate to their condition within the study. Treatment schools received free DIBELS and BURST services, while control schools received free DIBELS services. In the next chapter, we examine the extent to which schools "took up" these services. In this section, we discuss whether schools in the two arms of the analytic sample for the RCT differed on a number of pre-treatment covariates that prior education research suggests could be associated with the outcome measure used in this study—the *Star Early Literacy®* assessment.

Table 4.2 (page 16) compares control to treatment groups in the analytic sample terms of pre-treatment school characteristics. Once again, there are only small differences across treatment and control schools in the analytic sample, and a similar test used to test for statistically significant differences in the recruited sample was applied here to test for the aggregate of differences across treatment and control schools in the recruited sample on variables shown in Table 4.2 plus several other variables not shown in the table. That test also suggests no differences beyond what randomization alone would be expected to produce (χ^2=16.3 on 20 d.f.; p=0.7). So, it appears that processes of attrition did not disrupt the baseline equivalence of the analytics samples created by the original randomization procedure.

EQUIVALENCE OF TREATMENT AND CONTROL GROUP STUDENTS IN THE ANALYTIC SAMPLE

Table 4.2 compared treatment and control schools on a set of pre-treatment characteristics measured at the school-level. As we saw, the data suggested baseline equivalence on these characteristics, both for the full 92 school sample and 52 school analytic sample. We turn now to a comparison of treatment and control schools in the 52 school analytic sample in terms of various student-level characteristics. The relevant data for this analysis are shown in Table 4.3 (next page). In all, 26,907 unique students (about equally divided into treatment and control) participated in this study over the four-year data collection period, and Table 4.3 presents averages over students in each group on several student-level measures. One set is measures of students' baseline (i.e., pre-treatment) characteristics, the other are measures taken over the course of the study of student characteristics that should not have changed as a result of group assignment in the study. In the table, means are weighted consistent with study design. Specifically, means over the control group are calculated with weighs equal to a priori assignment to the treatment group, i.e. the same number that was factored in alongside of school size in the school-level comparisons mentioned above.

Differences in Baseline Characteristics: The first panel in Table 4.3 shows student characteristics measured at baseline. We begin by discussing students' DIBELS scores at the first time point in the study when they were tested with this instrument. As DIBELS was administered to nearly all students in both treatment and control schools, and because within treatment schools each student would have taken the DIBELS examination prior to receiving treatment, we can compare the two groups in terms of students' baseline DIBELS score. Differences along this measure were not large. Moreover, when this measure was combined with the other variables in the top panel (as well as additional variables), a combined difference test was not statistically significant (χ^2=6.9 on 11 d.f.; p=0.8).

2.2%, again in favor of the treatment group. The significance calculation cited above addresses pairing and clustering in the same manner as balance checks and is described below.

Table 4.3 | Comparison of Students in Treatment and Control School in the Analytic Sample on Selected Student-Level Baseline Measures

	Control Students	Treatment Students	Difference (T-C)	Standardized Difference
Characteristics at Entry:				
DIBELS Score at Join Point	104	108	4.2	0.041
Grade at Join Point	1	1	-0.034	-0.03
White	0.58	0.58	0.0053	0.011
Limited English Proficiency	0.1	0.14	0.036	0.11
Female	0.49	0.49	-0.0003	-0.001
Date of Birth	1/13/08	1/1/08	-12 days	-0.017
Eligible for Free or Reduced Lunch	0.66	0.73	0.066	0.144
Special Education	0.07	0.075	0.0051	0.02
Differences over Duration of Study				
Crossed Over	0.018	0.016	-0.002	-0.015
No. Times Retained in Grade	0.03	0.034	0.004	0.022
Ever below DIBELS benchmark	0.68	0.682	0.003	0.006

Differences in Other Characteristics. The second panel in Table 4.3 compares students on characteristics that should be *un*related to treatment but on which differences might emerge over the course of the study causing effects on the study's key outcome measure. The first variable is whether a student left the school to which he or she was initially assigned and joined another school in the study. Testing for differences in "crossover" is important given the design of the BURST efficacy trial because treatment and control schools were often co-located in the same district and the study took place over a four-year period. Given this, it is possible that crossover from treatment to control (or vice versa) occurred. Importantly, crossover *did* occur in the current study, although crossover rates were small and similar across treatment and control groups. Indeed, less than 2% of students in either group "crossed over" to the other group. There was also little difference in other variables which might affect student outcomes, such as whether students were retained in grade or ever scored below DIBELS time-of-year benchmarks. In fact, in a combined test, these differences were not statistically significant (χ^2=2.7 on 4 d.f.; p=0.6).

In analyses not shown here, we performed some additional (and more complex) balance tests. In these analyses, each student in the study was assigned a "join year" indicator, where join year was the year of the study that a student first appeared in any study school. These join year indicators were then interacted with student-level covariates including those shown in Table 4.3. Next, attention was restricted to the 10,400 students observed in the study only after the beginning of study year 1, either because they first joined the study later in year 1 or because they first matriculated at a study school in study year two or later. Then differences between treatment and control groups along these interaction variables were assessed using the combined differences procedure, again with adjustment for clustering by school and for the *ex-ante* pairing of schools. In a combined test, differences in these interactions were not statistically significant (χ^2=12.0 on 12 d.f.; p=0.4).

Summary: In summary, the analyses presented in this section suggest that despite high attrition rates at the recruitment stage of the study, the study team's policies of matching schools prior to random assignment and dropping all schools from any randomization cluster in which one school voluntarily withdrew resulted in roughly equal attrition rates across treatment and control groups and produced an analytic sample of treatment and control group schools that were not significantly different on a wide range of pre-treatment covariates. Moreover, balance on covariates at the start of the study persisted over all succeeding years of the study, with little evidence of differential crossover or selective enrollment in treatment schools.

Table 4.4 | Characteristics of BURST Efficacy Trial Schools

	Comparison Populations		RCT School Samples		
Variable	Population of U.S. Schools	BURST Subscriber Group	All Recruited Schools	RCT Attrition Sample	RCT Analytic Sample
District Characteristics					
Community Socio-Economic Status[3]	-0.03	-0.15	-0.52	-0.30	-0.68
Number of Students in District	16,366	45,202	8,315	8,191	8,410
School Characteristics					
Title I School (1=Title I)	0.80	0.78	0.97	0.95	0.98
Schoolwide Title I (1=Schoolwide)	0.54	0.61	0.88	0.77	0.96
Targeted Title (1=Targeted)	0.14	0.06	0.08	0.18	0.00
School in City (1= City)	0.28	0.36	0.37	0.53	0.25
School in Town/Rural (1=Town/Rural)	0.38	0.34	0.49	0.45	0.52
School in Suburb (1=Suburb)	0.33	0.31	0.14	0.03	0.23
Teacher - Child Ratio (z-scored)	0.07	0.06	0.06	0.07	0.06
Number of Students	468	466	401	419	387
Percent of White Students	54	48	50	38	59
Percent of Students on Free Reduced Lunch	55	58	73	70	75
Student Achievement					
Percent Students Proficient in Reading Grade 3 (z-scored)	0.03	-0.02	-0.21	-0.18	-0.23
Percent Students Proficient in Math Grade 3 (z-scored)	0.03	-0.01	-0.31	-0.15	-0.44
Missing Values					
Missing Teacher -Child Ratio (%)	03	00	00	00	00
Missing Student Pct. Proficient in 3rd Grade Reading (%))	16	11	18	25	13
Missing Student Pct. Proficient in 3rd Math (%)	16	11	16	25	10

THE ANALYTIC SAMPLE COMPARED TO TARGET POPULATIONS

This section turns from a discussion of baseline equivalence of treatment and control groups in the analytic sample to a detailed look at the kinds of schools included in the analytic sample—especially in comparison to populations that could be thought of as "targets for generalization" for the study's results. The two "target" populations discussed here are: (a) the population of U.S. schools that serve students in any of grades K-3; and (2) the population of 651 schools that taught students in any of grades K-3 and were active subscribers to Amplify's BURST services in AY 2016-2017.

Relevant data on this issues are shown in Table 4.4 (above). The columns in the table list the various groups of schools to be compared. The first column is the relevant US schools population, the second is the population of BURST users as of AY 2016-2017. For convenience, the table shows data for all schools recruited into the BURST efficacy trial, for the schools were dropped and retained from the study, and for treatment and control groups in the study. Rows in the table are selected characteristics of schools expressed as unweighted school means, where data in the table come from state education databases and are for the AY 2016-2017 school year.[7]

Recruited, Retained, and Analytic Samples: Recall that a goal of this study was to recruit schools with more than 50% of students eligible for the federal government's free and reduced price lunch program. The data in Table 4.4 show that this recruitment goal was achieved. For example, schools in the sample of 92 recruited schools averaged about 73% of students eligible for the lunch program, with 75% of students eligible in the analytic sample (compared to a US schools average of about 55% of students eligible). This study goal was achieved despite subtle differences in school locations and socio-economic composition between schools dropped from the study at the recruitment stage and those retained in the study for data collection and analysis. As Table 4.4 shows, schools in the analytic sample were more likely to be in districts located in towns or suburbs than dropped schools, to have a higher percentage of white students than dropped schools, serve fewer students, but also be located in communities with *less* advantaged socio-economic circumstances than dropped schools. However, rates of proficiency on state accountability assessments were roughly similar across the analytic and dropped samples.

Analytic Sample vs. the Relevant Population of US Schools: Many of these same differences are observed between schools in the analytic sample and the relevant population of U.S. schools (defined here as any regular public school serving students in any of grades K-3). Schools in the analytic sample are located in districts that are smaller than the average for schools in the relevant national population. Study schools are also more likely to be located in towns or a rural area than schools in the relevant national population, but the towns in which study schools are located are more socio-economically disadvantaged. Compared to the average school in the relevant national population, schools in the analytic sample also serve about 100 fewer students in grades K-3, have about 20% more students eligible for free or reduced price lunch, and have lower percentages of 3rd grade students scoring proficient or above on state assessments in the areas of reading and mathematics.

Analytic Sample vs. the Relevant BURST Subscriber Population: Table 4.4 also shows some differences between the schools included in the analytic sample and the relevant BURST subscriber population. As of AY 2016-2017, BURST subscriber schools were located in larger than average size districts, whereas the analytic sample comes from smaller than average size districts. Indeed, the difference in district size across the two groups is striking: 45,202 for the BURST subscriber population versus 8,410 for the analytic sample. Schools in the analytic sample are also located in less advantaged towns than schools in the BURST subscriber group, and these schools serve fewer students. Schools in analytic sample serve proportionally more White students than schools in the relevant BURST subscriber population, but these students more likely to be eligible for free or reduced price lunch, and less likely to be scoring at or above proficiency on state assessments than students attending schools in the relevant BURST subscriber population. Finally, all but one of the schools in the analytic sample (96%) operate a school-wide Title I program, whereas 61% of schools in the relevant BURST subscriber population and 54% of schools in the relevant US population operate such programs. In a school wide program, all students (not just identified students) are eligible to receive Title I funded services.

SUMMARY

This chapter described the processes of recruitment, random assignment, and attrition that produced the analytic sample for this study. Although there was considerable attrition of schools from the study during the

[7] State data for AY2016-2017 were obtained from data provided by SchoolDigger.com and by the Stanford Education Data Archive.

recruitment phase, the policies of the study team assured that attrition was roughly equal across schools assigned to treatment and control conditions. For that reason, schools in the analytic sample were balanced on covariates measured prior to launch of the study in AY 2013-2014. Equally important, balance in student characteristics was maintained across study years. Attrition processes did, however, alter the composition of the study sample, as schools from bigger and more urban districts were more likely to be dropped from the study, and this resulted in a final analytic sample of schools that was more likely than relevant "target" populations of inferences to be in a socio-economically disadvantaged small town, serving higher percentages of white students eligible for free or reduced price lunch, who were somewhat less likely to score at or above proficiency levels on 3rd grade state assessments in reading and mathematics.

5 | IMPLEMENTATION

This section discusses findings on program implementation in schools participating in the BURST efficacy trial. Two features of this trial make implementation processes an important topic of investigation. To begin, the BURST efficacy trail involved an experimental manipulation in which both treatment *and* control schools were offered free services from Amplify. In particular, control group schools were offered free access to (and training and support for) Amplify's *mClass: DIBELS Next*® formative assessment services, while treatment group schools were offered free training and support for *mClass: DIBELS Next*® plus *BURST: Reading*®. Importantly, the provision of DIBELS services to control schools means that the control condition in the current study is *not* a "business as usual" condition. It was instead a "DIBELS only" condition, and for this reason, we here examine how DIBELS was taken up in control group schools and how that compared to the ways in which DIBELS was taken up in control schools. Beyond this issue, however, a second reason to be interested in implementation in the current study is that Amplify did not require treatment schools to implement the BURST program uniformly. Rather, as discussed in Section 2 of this report, treatment schools were given considerable discretion to implement BURST in ways that suited their local contexts, just as they would have been had they purchased BURST in a market transaction with Amplify. For this reason, we might expect BURST program implementation in study schools to vary along a number of dimensions, including how much training and support treatment schools requested from Amplify, the personnel schools used to manage and deliver BURST instructional services, the percentage of students actually placed into BURST instructional groups, and the actual scheduling of BURST instruction for students. Thus, potential for variability in BURST implementation is another reason why questions about implementation are important in this study and why, earlier in this report, we posed a number of research questions about BURST implementation in study schools.

DIBELS SERVICE DELIVERY

We begin our discussion of implementation with an analysis of data on schools' uptake of the DIBELS assessment services provided by Amplify as part of the study protocol. Table 5.1 shows the relevant data. Over the four-year course of the study, there was uptake of Amplify's offer of DIBELS services in both treatment and control schools. For example, Table 5.1 (next page) shows that 22 of 25 (88%) control schools and 24 of 27(88%) treatment schools received at least one DIBELS training session from Amplify. Table 5.1 does show, however, that control schools received about one more DIBELS training over the four-year period than did treatment schools (2.68 trainings on average for control schools vs. 1.68 trainings for treatment schools), and this difference is statistically significant (t = -3.95, p = .000). Despite this difference in uptake of training, however, the data in Table 5.1 show that control and treatment schools conducted DIBELS assessments regularly over the four-year period of the study. In control schools, for example, 87% of all students were assessed at the beginning of the year and 92.5% were assessed in the middle of the year, while in treatment schools, those percentages were 89.8% for the beginning of the year and 95% for the middle of the year. This small difference was not statistically significant. Thus, while the average treatment school received about 1 more DIBELS training over the course of four years than did the average treatment school, rates of DIBELS assessment were similar at treatment and control schools over the course of the four-year study.[8]

[8] In year four of the study, we administered a survey to teachers in treatment and control schools to examine whether the professional development experienced by teachers (outside of that provided by Amplify) differed systematically

Table 5.1 | DIBELS Training and Use In Control vs. Treatment Schools

Implementation Measure	Control Schools (n=25)	Treatment Schools (n=27)
Number of Schools Receiving *Any* DIBELS Training or Support from Amplify	22	24
Number of DIBELS Trainings Per School	2.68	1.68
Percentage of Students Assessed with DIBELS (Beginning of Year)	87.3%	89.8%
Percentage of Students Assessed with DIBELS (Middle of Year)	92.5%	95.1%

Table 5.2 | BURST Services Delivered to Treatment Schools

Service Type	Treatment Schools Mean Number of Visits
Any Service	6.8
BURST Training	2.2
BURST Support Visit	4.5

BURST SERVICE DELIVERY IN TREATMENT SCHOOLS

In addition to the Amplify offer of free training for DIBELS use to both treatment and control schools, Amplify offered treatment schools free training and support services for BURST throughout the four-year period of the study. Amplify's service delivery records show that 26 of the 27 schools in the treatment group availed themselves of this offer—with one treatment school not taking up this offer. Table 5.2 (above) shows some relevant statistics about these services. On average, schools in the treatment group were visited by Amplify staff on 6.8 occasions over the four year period of the study (with a range of 0 to 11 visits per school). On average, 2.2 of these visits were for training in the use of BURST (range = 0 -5), while 4.5 visits were in support of BURST use (range = 0-8). Most schools received one training early in the study and another at some later point, while support visits occurred over the course of the study.

ORGANIZATION OF BURST PROGRAM AT TREATMENT SCHOOLS

Over all four years of the study, we also charted how schools organized to assign students to BURST groups and deliver BURST instruction. This work showed that two treatment schools (in the same district) *never* organized systematically to provide BURST instruction to students (although at least one teacher in each school *did* provide BURST instruction to students over the course of the study). In addition, another three schools (in a single district) stopped systematically organizing for delivery of BURST instruction in the third and fourth years of the study.

Table 5.3 (next page) shows the relevant data on how schools organized to deliver BURST instruction for students. In the three schools that were organized for just two years and in the remaining schools 22 treatment schools, we observed varying patterns of organization for BURST implementation. In 18 of 25 schools, literacy coaches handled the assignment of students to BURST groups, while in 6 schools this was handled by classroom teachers, and in one school it was handled by the principal. In 13 of 25 schools, classroom teachers provided BURST instruction, while in the 12 other schools where BURST was systematically organized,

across treatment and control schools. The results of that analysis (shown in Appendix C) showed no statistically significant differences in the types or amounts of reading-related professional development experienced by treatment vs. control teachers in year four of the study.

Table 5.3 | Organization of BURST Program in Treatment Schools

Organizational Form	Number of treatment schools
No systematic organization	
All four years	2
Last two Years	3
Who Assigns Students to BURST	
Classroom Teacher	6
School Literacy Coach	18
Principal	1
Not recorded	2
Who Provides BURST Instruction	
Classroom Teacher	13
School Literacy Coach	12
Not recorded	2
Schedule for BURST Instruction	
Literacy Block	15
Intervention Block	10
Not recorded	2

literacy coaches provided BURST instruction. In 15 of the 25 schools where BURST was systematically organized, BURST instruction was provided during the regular literacy block while in the other 12 schools, BURST instruction was provided during an intervention block. Overall, nine of 25 schools had literacy coaches assign students to BURST groups and offer instruction in an intervention block, five schools had literacy coaches assign students to BURST groups and provide instruction in the regular literacy block, another five schools had literacy coaches assign students to BURST groups but had classroom teachers provide instruction during the regular literacy block, four schools had classroom teachers assign students to BURST groups and deliver instruction in the regular literacy block, and another four had classroom teachers assign students to BURST groups and deliver instruction during an intervention block. Thus, as expected, there were varying patterns in how treatment schools organized to provide BURST instruction.

PROVISION OF BURST INSTRUCTION TO STUDENTS IN TREATMENT SCHOOLS

We turn now to the question of how much BURST instruction was received by students in treatment schools. To address this question, University of Michigan used data gathered from Amplify's *mClass* and BURST data systems during each of the four years of the study. In any given year of the study, we used the *mClass* data system to obtain records of each student's DIBELS score at the beginning, middle, and end of year assessment points, and in what follows, we classified these scores into three groups at any given assessment point during a school year: "Red" if a student's score was well below the DIBELS grade-level/time of year benchmark for a time point, "Yellow" if a student's DIBELS score was below (but not well-below) the relevant grade-level/time of year benchmark for a time point, and "Green" if a student's DIBELS score was at or above the relevant grade-level/time of year benchmark. In treatment schools, we then used the BURST data system to record the number of 10-day BURST cycles of instruction received by a student in Fall and Spring semesters, again doing so at each of the four years of the study. When combined, these data allow us to describe whether nor not a student who was classified as Red, Yellow, or Green in a given semester received BURST instruction in that semester and how many cycles of BURST instruction were received by that student during that semester. Overall, when data from semesters are added together to get an annual observation for each student in any given year of the study, we have 26,813 distinct annual records for 14,165 unique students who spent at least one year during the study in one of the 27 treatment schools under study.

Table 5.4 | Percentage of Students Receiving Any BURST Instruction by Year, Semester, and DIBELS Status

	Red		Yellow		Green	
	Semester 1	Semester 2	Semester 1	Semester 2	Semester 1	Semester 2
Study Year 1	56%	53%	53%	53%	13%	14%
Study Year 2	45%	45%	43%	39%	11%	10%
Study Year 3	39%	38%	40%	39%	10%	10%
Study Year 4	37%	34%	33%	34%	10%	7%

Percentage of Students Classified as Green, Yellow, or Red Receiving BURST Instruction by Semester and Year of Study: The initial question we asked was about the percentage of students who actually received BURST instruction in any given semester of the study. Those data are reported in Table 5.4 (next page). As the table shows, schools did not serve *all* students, nor did they serve all students for whom the BURST program is intended (i.e., students scoring Red or Yellow on DIBELS). Instead, schools tended to give about equal priority to students classified as Red or Yellow in a given semester and to serve students classified as Green at much lower rates. There is also a noticeable decline in service rates across years of the study. In year one of the study, for example, a little over 50% of students classified as Red or Yellow received BURST instruction and about 13% of students classified as Green received BURST instruction. By year four of the study, those numbers were around 35% for Red and Yellow students and 10% for Green students.

Variability Within and Between Schools in Amount of BURST Instruction Received: The data in Table 5.4 suggest that schools were not able to serve all struggling readers, and in analyses not shown here, we further found that schools often served many students only for a single semester in a given year. Because of this, we next asked about the amount of service students received across a year (not a semester). Figure 5.1 (next page) shows some initial data relevant to this question. The figure displays a histogram for each school in the study showing the distribution in number of BURST cycles per year for students (averaged across all years of the study). What this figure shows is that the largest number of students in a school always receive *no* BURST instruction, but that among students who were served, there is a wide distribution of number of BURST cycles received, both within the same school, and across schools.

To better model this phenomenon, we developed two, related statistical models. Both models use the same variables to predict the number of cycles of BURST instruction a student received in a given year, but one model includes only students who received BURST instruction in a year and the other is for *all* students. Both are straightforward linear mixed models in which the number of cycles (y) a student (i) in school (j) received per year was seen as a function of whether or not the student had scored at the Red/Yellow level on DIBELS at any point in a year, the year of the study (year = 1, 2, 3, or 4), a student random effect, and a school random effect. The results are shown in Table 5.5 (page 27). In these models, the intercept is an estimate of the average number of BURST cycles received by students who were classified as Red/Yellow in a year, there is a year effect, and one can obtain an estimate of the number of cycles received by a student who was always classified as Green by subtracting the always Green coefficient from the intercept. The variance components are also relevant here. The variance between schools can be used to understand mean differences in service levels across schools, the student-level variance can be used to understand variance in services received among students within the same school, and the residual reflects variance in the number of received BURST cycles across years of observation and other sources of prediction error.

Let us begin our discussion of these statistical models by looking at left hand side of Table 5.5, which displays the results for *served* students. The data here show that the average student who was classified as Red/Yellow during a given year and who also received BURST instruction in that year experienced 5.94 cycles of BURST instruction—about what Amplify recommends for a semester of service, but half what the company recommends for a full year. Importantly, the variance components suggest that this service level does not vary among students in the same school (variance = 0), but the amount of instruction received does vary: (a)

Figure 5.1 | Histograms of BURST Cycles Per Year Received by Students in the 27 Treatment Schools

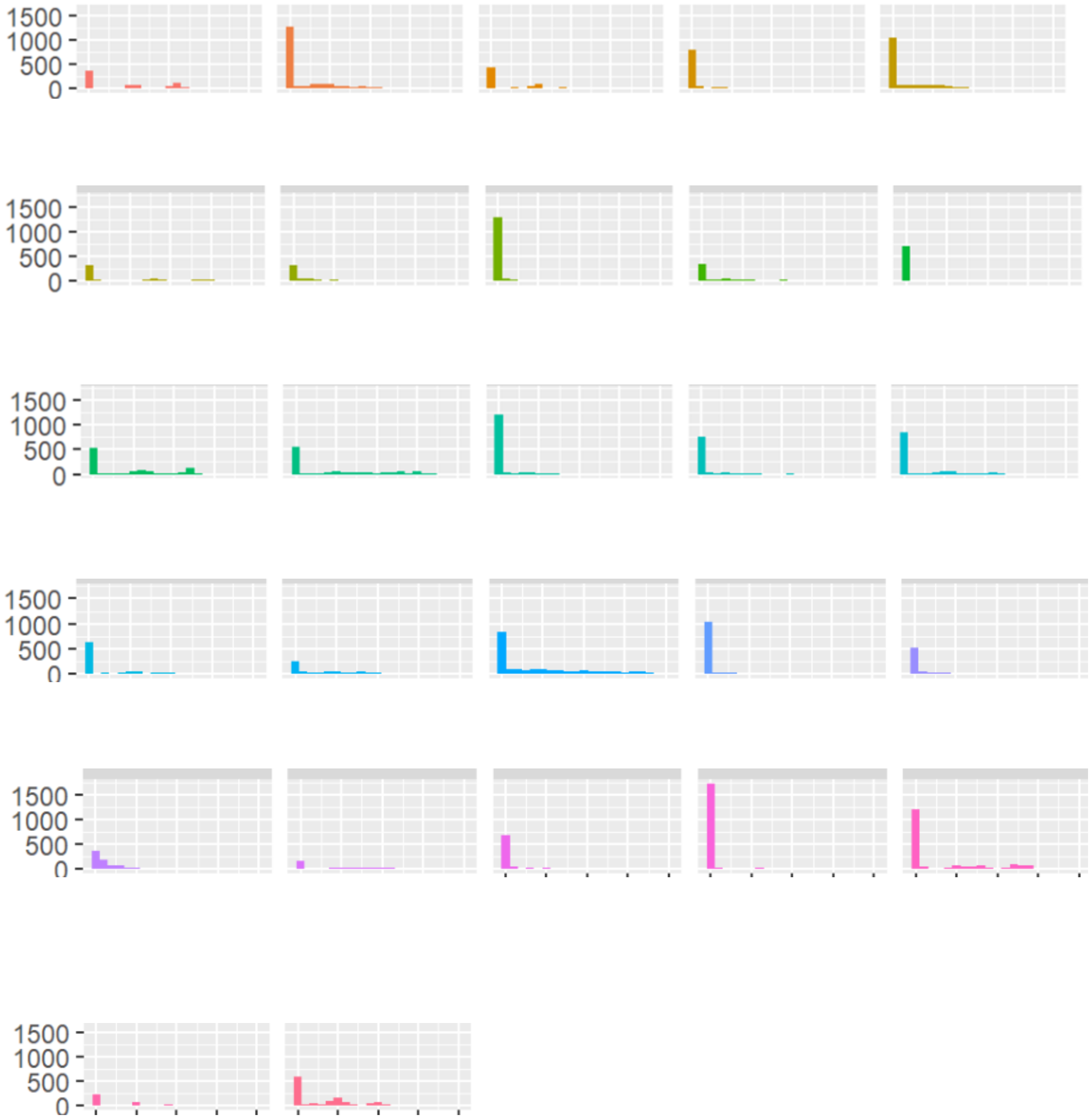


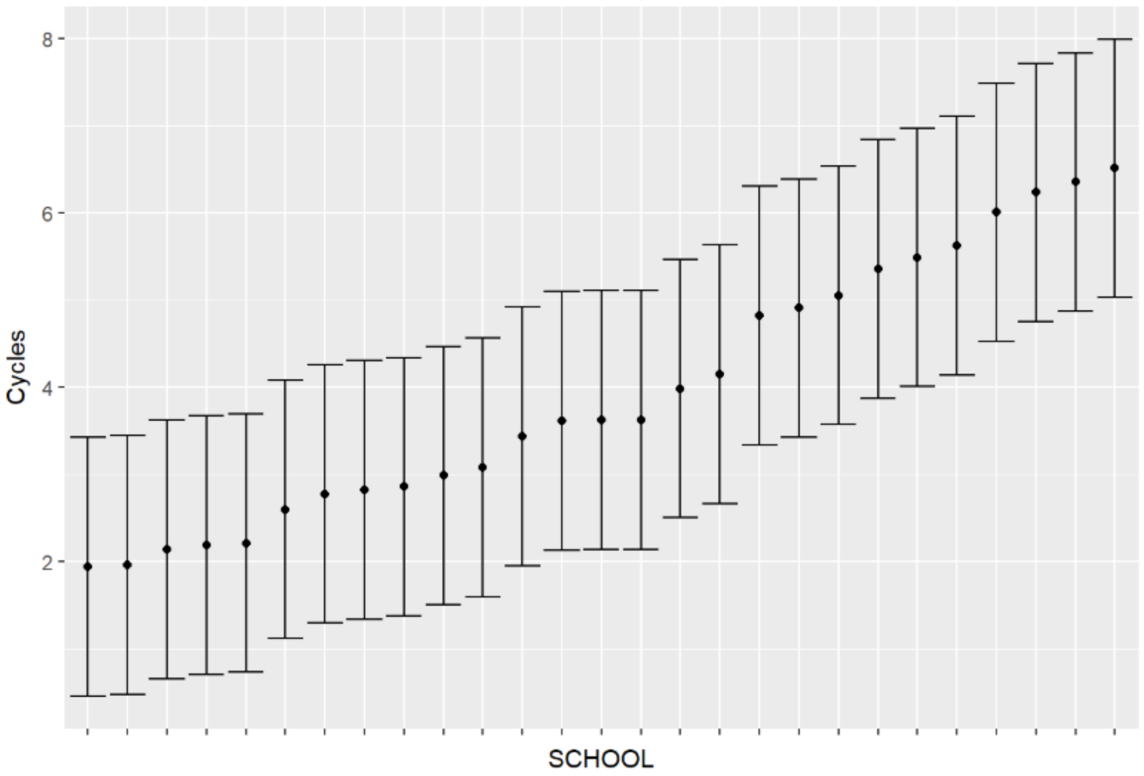
Table 5.5| Results from a Linear Mixed Effects Model Predicting Number of Cycles of BURST Instruction Received by Students as a Function of Study Year, DIBELS Status, and Student and School Random Effects

	Served Students			All Students		
	Coefficient	Standard Error	T-value	Coefficient	Standard Error	T-value
Intercept	5.94	.49	13.23	4.15	.29	14.30
Study Year	.015	.033	.461	-.22	.02	-13.97
Student Always Green	-2.22	.077	-28.72	-2.82	.04	-74.44
	Variance Components					
Error	9.16			6.81		
Students	0			1.28		
Schools	5.22			2.21		

across years (residual variance = 9.16) and among schools (variance = 5.22). Meanwhile, in the model for all students, we can see that any student who was classified as Red/Yellow in a given year is expected to receive 4.15 cycles of BURST instruction. This reflects the fact that not all students classified as Red/Yellow received BURST instruction in a given year. In this statistical model, there is considerable variance in service receipt—both across schools (variance = 2.21) and among students in the same school (variance = 1.28).

Of the two statistical models in Table 5.5, the "all students" model provides the most relevant metric for measuring the amount of BURST instruction delivered by schools to struggling readers (i.e., students classified as Red/Yellow on DIBELS). As we have seen, in any given year, schools do not serve all students identified as struggling readers, and when they do, schools vary in in the number of cycles of BURST instruction they provide. The intercept in the "all students" model takes this into account. It is the average level of a BURST instruction a Red/Yellow student can expect to receive in a given year. Figure 5.1 (next page) provides a visual representation of the results from the all students model. Each dot in the figure represents the random effect of each treatment school in the study, where the dot is the mean number of cycles of BURST instruction per year predicted for Red/Yellow students from the linear mixed model. The error bar around each dot shows the variation in cycles of BURST instruction among students *within* the same school as predicted by the linear mixed model, where this error bar is set at +/- one standard deviation calculated from the within school variance. The assumption in the model (and the figure) is that within school variance is the same in all schools. Looking at the figure, the reader will see that the average number of BURST cycles received by students ever classified as Red/Yellow in a given year ranged from a low of about two cycles per year in the two schools at the far left of the figure to a high of about 7 cycles per year in two schools at the far right of the cycle. Within a given school, the variance component for students shown in Table 5.4 (and the error bars in Figure 5.1) suggest that students at the same school are not likely to receive equal numbers of BURST cycles, even after controlling for DIBELS status at entry. In fact, from the statistical model for all students in Table 5.5, we can see that within the average treatment school, 68% of Red/Yellow students will receive between 3 and 5.25 cycles per year in schools. So, there is variation in the expected number of BURST cycles per year received by struggling readers, both within and between schools.

Accumulation of BURST Instruction across Years of Exposure to Treatment: The data discussed so far described the expected number of BURST cycles of instruction a given student would be expected to receive in a given year (conditional on DIBELS status). This is important data, but many students will stay in a BURST school over multiple years, and in that case, a student might receive BURST instruction in more than one year. Figure 5.2 (page 28) shows the rate at which students accumulate BURST instruction across multiple years of exposure to treatment. In the figure, students are classified as Red, Yellow, or Green based on their DIBELS assessment score *at entry* into the study. The figure then shows the expected number of BURST instructional cycles these different groups of students would be expected to receive at the end of one, Figure 5.1| Graph Showing Model-Based Estimates of Within and Between School Variation in the Average Number of Cycles Per Year of BURST Instruction Expected for Red/Yellow Students By Treatment School **



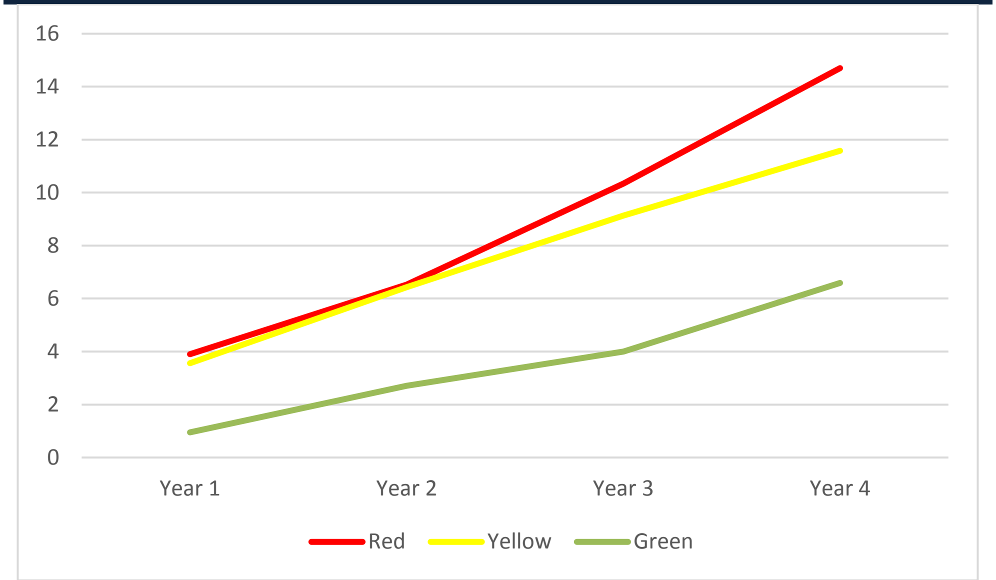
Each dot represents the ***estimate* for average number of BURST cycles received by Red/Yellow students per year at a school. Error bars are the *within-school* standard deviation of cycles received as estimated from the model under the assumption that this is the same in all schools. Plots are based on results from the "all students" model shown in Table 5.5.

two, three, and four years of attendance at a treatment school in the study. Recalling that 26 of the 27 treatment schools operate school-wide Title I programs, we see that even students who entered a treatment school at or above DIBELS benchmarks can be expected to receive some BURST instruction as they continue in a treatment school, but much less than students who entered below and far below DIBELS benchmarks.

BURST Instruction Received by Students in Treatment Schools vs. Amplify Recommendations

It is now time to discuss the extent to which treatment schools in the BURST efficacy trial delivered BURST instruction at the "ideal" level recommended by Amplify. We have just seen (in Table 5.5) that the average student scoring Red or Yellow in a given year is expected to receive less than the Amplify-recommended 12 cycles (or 120 days) of BURST instruction per year. Instead, Table 5.5, showed that a student classified as Red/Yellow in a given year and who actually received BURST instruction was expected to get about 6 cycles of BURST instruction. However, since not all Red/Yellow students in a school actually received BURST instruction, a better metric for understanding how much BURST instruction struggling readers got in BURST treatment schools is to look at the average number of BURST cycles any student classified as Red/Yellow

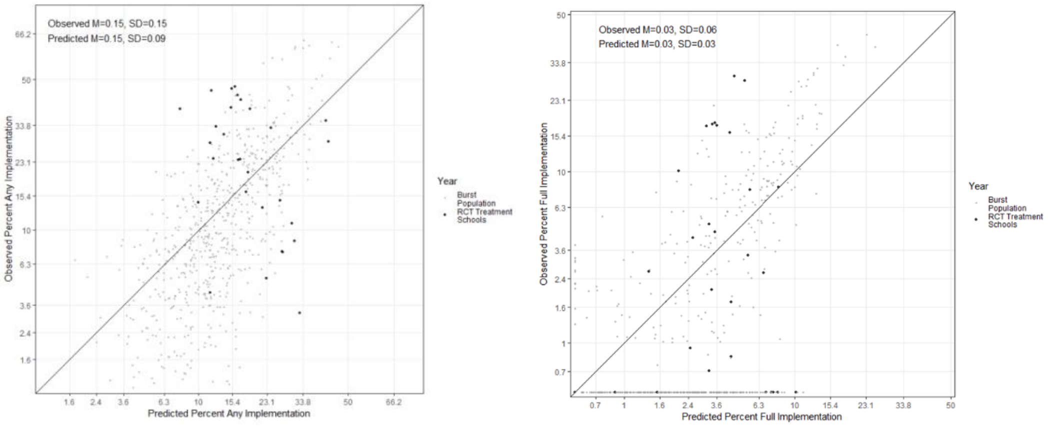
Figure 5.2: Cumulative Number of BURST Cycles Across Years of Exposure to Treatment for Students Who Entered Study as Red, Yellow, or Green on DIBELS



was expected to receive in a year of the study. From Table 5.5 (right hand panel), we estimated that Red/Yellow students averaged about four cycles (or 40 hours) of BURST instruction per year. This is far lower than the 12 cycles (or 120 hours) per year of BURST instruction that Amplify recommends for struggling readers. In fact, as Figure 5.2 (above) shows, the average struggling reader in this study would not be expected to receive this level of BURST instruction until that student had been in a BURST school for four continuous years. So clearly, on the basis of the data presented here, the average struggling reader in a treatment school in this study did not receive an "ideal" dose of BURST instruction in a given year.

It is possible, however, that the levels of instruction provided to struggling readers in this study are about what would be expected under "routine" conditions to implementation. To explore this issue, we acquired data from Amplify on the 651 schools serving any of grades K-3 who also subscribed to *BURST®: Reading* in AY 2016-2017. The characteristics of these schools and the differences between this group and treatment schools were described earlier (in Section 4) of this report. In this section, we compare two pieces of data on BURST instruction that were present in the data we received from Amplify and that were also collected on treatment schools in the study. The first piece of data was the percentage of students in a school who received *any* BURST instruction in a given semester, which we averaged across grades and semesters for a given school. The average is only for AY 2016-2017 for schools in the BURST user population, but the average for treatment group schools is taken across all four years of the study. In the BURST user population, 15% of all students received at least 1 cycle of BURST instruction in AY 2016-2017 (s.d. =15%), while in treatment schools, that percentage was 25% of all students (s.d.= 15%). So, treatment group schools provided BURST instruction to proportionally *more* students each semester than did the AY 2016-2017 BURST user group. However, none of the schools provided many students with what Amplify would consider to be the "ideal" number of BURST cycles. In the BURST user population, the percentage of students receiving 12 cycles in AY 2016-2017 was about 2.5%, while in the treatment group schools in this study, that number was 3%. Again, the treatment schools offered 12 or more cycles of BURST instruction to proportionally *more* students each year than did the AY 2016-2017 BURST user group, although in both groups, the proportion is very small.

Figure 5.3: Predicted Versus Observed Provision of BURST Instruction to Students in the 27 Treatment Schools Study



Care should be taken, however, before concluding that treatment schools in the current study provided more BURST instruction to students than would be expected under "routine" conditions of implementation. A problem with that inference is that treatment group schools in the BURST efficacy trial differ in important ways from the AY 2016-2017 user group on several school-level characteristics that might predict BURST instructional provision at a school. To test this idea, we used a Bayesian Adaptive Regression Tree (BART) model as discussed in Kapelner & Bleich (2013) to create an "implementation prognostic score" for treatment schools using data on the AY 2016-2017 BURST user group. In this approach, a variety of school and district demographic variables (described in Appendix B) were used to predict the percentage of students in the user group schools that received any treatment. Once that model was fit, we applied it to the treatment group schools in order to get the model-predicted percentage of students expected to receive any BURST instruction given treatment school demographics. Figure 5.3 (above) shows the results of that analysis. In the left hand graph, the X (or horizontal) axis is the model-predicted percentage of students receiving *any* BURST instruction in a given year while the Y (or vertical) axis is the observed percentage (where the axes are on a logit scale for convenience of presentation). Treatment schools are denoted by a black dot, whereas AY 2016-2017 users are denoted by a gray dot. Black dots above the reference line had a greater observed than predicted percentage of students receiving BURST instruction, while schools below had lower than expected provision. The left hand graph shows two things. First, treatment schools tended be at the higher end of predicted implementation, as shown by the location of black dots toward the right hand side of the graph. Second, 16 of 27 treatment schools were observed to provide more instruction than predicted by the BART model while 11 others showed lower than predicted observed values. A cautious interpretation of these data is that the treatment group—as a whole—provided about the same amount of BURST instruction to students as would be expected under "routine" conditions of implementation (although some provided more than expected, and some provided less). The graph on the right (above) is formatted in the same way as the graph to the left, but in this graph the X axis is the predicted percentage of students in a school who get a "full" set of 12 BURST cycles in a year. Note that a few treatment schools and many schools in the user population have an *observed* percentage that is at or near zero. Moreover, note that treatment schools in the right hand graph are in the lower two-thirds of predicted implementation. Still, the overall conclusion from this graph is the same as the conclusion from the left hand graph. As a group, schools in the treatment group provided "full" amounts of BURST instruction at rates that approximate what would be expected under "routine" conditions of implementation (with some providing more than predicted and others providing less).

OBSERVATIONS OF BURST INSTRUCTION

As a final check on BURST implementation in treatment schools, we had Amplify professional services and research staff use a fidelity of implementation checklist to observe the delivery of BURST instruction in 25 of the 27 schools. The checklist had observers record a number of features of each BURST lesson they observed. We focus here on the low inference items on that checklist. Each observer recorded the day of the BURST cycle being observed (day=1-10), the length of the BURST lesson observed, the extent to which materials for the lesson were present, and the extent to which the teacher followed the order of activities listed in the lesson script. These last two variables were coded as low (1), medium (2), or high (3). Overall, 6 different observers observed 382 BURST lessons across the schools—an average of 16 observations per school (range = 2 to 33). The average BURST group size observed was 4.34 students (s.d. = 1.8), and the average day of a BURST cycle observed was 2.84 (s.d. = 1.48). Across all lessons, teachers were observed using BURST materials (mean = 2.78, with 83% of lessons scored as 3). Also, teachers overwhelmingly followed the order of the lesson script (mean = 2.83 with 86% of lessons scored as 3). Thus, the data suggest that the major difference in program implementation across schools was not *how* teachers acted when delivering instruction in BURST groups, but rather the amount of BURST instruction students received in different schools.

SUMMARY

This section reported data on the extent to which treatment and control schools took up Amplify's offer of free training and services and how treatment schools organized the delivery of BURST instruction to students. The data showed that both treatment and control schools tested students with DIBELS at the same rates (≅ 98%). But there was strong evidence that treatment schools differed in how they organized and delivered BURST instruction. Two treatment schools never organized systematically to deliver BURST instruction and served very few students. The remaining 25 treatment schools organized the BURST program differently, assigning different personnel the roles of grouping students for BURST instruction and delivering instruction. Schools also differed in whether they delivered BURST instruction during the regular reading period or an intervention period. Overall, treatment schools began the study by delivering BURST instruction to about 50% of their struggling readers, but this percentage declined across years of the study. There also was a great deal of variation across schools in how much instruction the average "struggling reader" received. In the average school in the study, a student classified as Red or Yellow on DIBELS could expect to receive just over 4 cycles (or 50 hours) of BURST instruction in a given year. While this is below Amplify's recommended provision of 12 cycles per year, it is about what would be expected under "routine" conditions of implementation.

6 | ANALYSIS OF BURST PROGRAM EFFECTS ON STUDENT ACHIEVEMENT OUTCOMES

This section describes how we formally tested a series of hypotheses about the effects on students' early reading achievement of attending a school assigned to the BURST treatment. To do so, we briefly review the study design, describe the student achievement measure that serves as the study outcome, list the formal hypotheses to be tested, and describe our strategy for estimating treatment effects. We then present the results of a set of statistical tests of hypotheses about BURST treatment effects on students early literacy learning.

STUDY DESIGN

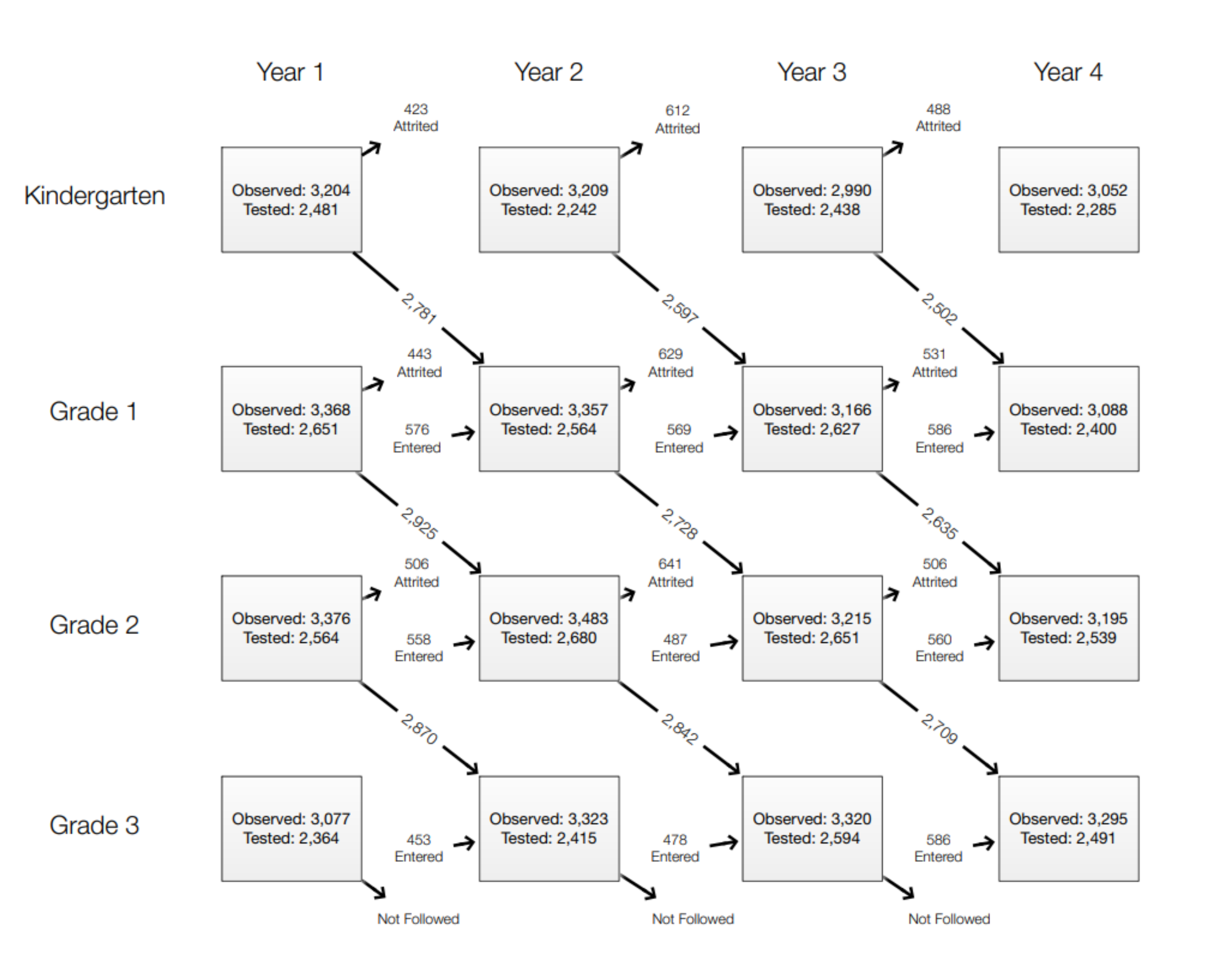
Treatment and Control Conditions. The reader will recall that the BURST efficacy trial was conducted as a cluster randomized field trial in which two treatments were randomly assigned to schools using procedures described in Section 4 of this report. The treatment of interest in the study was free access to services and training for use of Amplify's *BURST®: Reading* program, a supplemental reading program intended for use with struggling readers (as described in Section 2 of this report). The other treatment, which we call the control condition, was free access to Amplify's *mCLASS: DIBELS Next®* assessment services and training opportunities.

As discussed in Sections 2 and 5 of this report, Amplify did not require schools to implement either DIBELS or BURST in a standardized fashion, and so the efficacy trial discussed here took the form of what is sometimes called an "encouragement" design, where the term is used to denote that the experimental manipulation takes the form of *encouragement* to use a treatment, not a supervised and uniform application of a treatment. As discussed in Section 5 of this report, both treatment and control schools implemented treatments that involved use of DIBELS testing, where treatment and control schools administered DIBELS assessments at similarly high rates). Among BURST treatment schools, however, there was variation in how much schools took advantage of BURST training and support services, in how treatment schools organized to manage and deliver BURST instruction, and in the average number of BURST instructional cycles offered to students. On average, however, we showed in Section 5 of this report that treatment schools tended to provide about the amount of BURST instruction that would be expected to be delivered to students had these schools been implementing the program under "routine" conditions of implementation.

Student Observations. The study was launched in AY 2013-2014 when Amplify began offering free services to treatment and control schools. Free service were then offered for four consecutive years, ending in 2016-2017. During this time period, the research team gathered data on the administration of DIBELS tests to all students in grades K-3 in control schools, and the study gathered data on provision of BURST services to all students in grades K-3 in treatment schools. Any student enrolled at a study school in grades K-3 during the entire study period was considered eligible for data collection, and data were collected on students continously over the study period. In all, data were gathered on 26,907 unique students, with the average student contributing 1.89 observations to the study. Of the 26,907 students in the study, 13,572 were in treatment schools only (and contributed an average of 1.92 observations per student), 13,969 students were in control schools only (and contributed an average of 1.87 observations per student), and 634 students crossed over from treatment to control or vice versa and were assigned for analysis purposes to treatment or control conditions depending on the assigned status of the school where they were first observed.

Figure 6.1 (next page) shows how the student sample developed over time and why the expected number of observations per unique student naturally varies across students, both as a function of attrition from the study and as a function of when a student first entered the study. The figure also shows the number of students

Figure 6.1 | The BURST Effiacy Trial Student Sample By Year of Study


Year 1
Year 2
Year 3
Year 4
Kindergarten
Grade 1
Grade 2
Grade 3
Observed: 3,204 Tested: 2,481
423 Attrited
Observed: 3,209 Tested: 2,242
612 Attrited
Observed: 2,990 Tested: 2,438
488 Attrited
Observed: 3,052 Tested: 2,285
2,781
2,597
2,502
Observed: 3,368 Tested: 2,651
443 Attrited
576 Entered
Observed: 3,357 Tested: 2,564
629 Attrited
569 Entered
Observed: 3,166 Tested: 2,627
531 Attrited
586 Entered
Observed: 3,088 Tested: 2,400
2,925
2,728
2,635
Observed: 3,376 Tested: 2,564
506 Attrited
558 Entered
Observed: 3,483 Tested: 2,680
641 Attrited
487 Entered
Observed: 3,215 Tested: 2,651
506 Attrited
560 Entered
Observed: 3,195 Tested: 2,539
2,870
2,842
2,709
Observed: 3,077 Tested: 2,364
453 Entered
Observed: 3,323 Tested: 2,415
478 Entered
Observed: 3,320 Tested: 2,594
586 Entered
Observed: 3,295 Tested: 2,491
Not Followed
Not Followed
Not Followed

enrolled at all schools in a given year (and thus eligible for outcomes data collection) and the number of students on whom outcome data were actually collected that year.

Figure 6.1 shows quite clearly that we have repeated observations on (most) students, but the number of observations per student varies as a result of the study design. Moreover, the reader can easily see why, as a result of the study design, we have more data points on students after one and two years of followup than three and four years of followup. Indeed, about 50% of all student data come after just one year of followup, another 30% come after two years of followup, 16% come after 3 years of followup, and only 4% come after 4 years of followup. The reader will also note from Figure 6.1 that about 77% of enrolled students were assessed on the outcome measure in any given year. However, no effort was made in this study to impute missing outcomes data for students not assessed, in part because the data on differential attrition shown in Table 4.3 of this report did not show a pattern of differential attrition across treatment and control students. As a result, analyses presented below are conducted only on those students for whom outcome data is available.

THE OUTCOME MEASURE

Outcome assessments were adminsistered at school sites during a specified Spring testing window each year. In most cases, schools were responsible for administration of the outcome assessment, but in some schools (where compliance to the study's test administration protocol proved difficult to achieve), Amplify staff managed the assessment process. The outcome assessment used in this study was *STAR Early Literacy™*, a standardized assessment developed and sold by Rennaisance Learning, Inc. A complete description of this assessment can be found in the *STAR Early Literacy™ Technical Manual* (available here: http://doc.renlearn.com/kmnet/ r004384710gj119f.pdf).

Test Content. *STAR Early Literacy™* (SEL) was developed for use with students in grades K-3 and is used to assess student knowledge in 11 areas, including: the alphabetic principle, concept of word, visual discrimination, phonemic awareness, phonics, structural analysis, vocabulary, sentence-level comprehension, paragraph-level comprehension, and early numeracy. Each of these areas is further defined by component skill sets (e.g., phonemic awareness is defined by 11 skill sets such as blending phonemes and phoneme segmentation), and each skill set is further decomposed into discrete skills around which item banks have been developed (e.g., phoneme segmentation is defined by segmenting syllables in single-syllable words and segmenting syllables in multi-syllable words). The content areas tested by SEL cover four of the five key skills identified by the National Reading Panel (NRP) as being associated with successfully learning to read in the early stages of literacy acquisition. The only NRP skill domain not assessed by SEL is reading fluency.

Test Administration. SEL is a computer adaptive assessment. Students interface with a computer to take the test, and test software provides students with the verbal instructions and visual input need to understand and respond to each question so a proctor is not needed to read questions aloud to students or record student repsonses. Items are administered to students through a process of adaptive branching, a process that gives items to students based on their previous item responses such that students generally answer items at a rate of about 70% correct. A test session includes instructions for taking the test as well as administration of 27 items per student. Items have time limits established using latency data from a large calibration sample, and students who time out on an item are scored as having an incorrect response. Test sessions are expected to be 8 to 15 minutes in length (including test instructions). As discussed above, testing for this study was conducted during a common Spring time period across years of the study. In most cases, Amplify notified schools of the testing period, allowed schools to administer the assessments, and then monitored test completion rates, although in a few schools, test administration was directly conducted by Amplify researchers. Student test-taking rates by year and grade varied between a low of 70% of enrolled students in grade K taking the assessment in year two of the study to a high 83% of eligible students taking the test in grade 1 of the third year of the study, with a median of 77% of eligible students taking the outcomes assessment at any grade/ year of the study.

Use of SEL Scale Scores as Outcome Measure. The SEL scale score provided by the test publisher for a given student in a given year was used as the outcome measure in this study. The scale score is a global measure of a student's early reading ability based on the θ obtained from the publishers application of Item Response Theory to the data. The raw θ's can range from -6.00 to +6.00 and scale scores will range from 300 to 900.[9]

Psychometric Properties of SEL. Based on calibration sample data, the publishers report a split half reliability of .91, and a "generic" reliability .92 (where generic reliability is defined as $[1 - (\sigma^2_{\text{error}}/\sigma^2_{\text{total}})]$. In this formulation, σ^2_{total} is the total variance in test scores and σ^2_{error} is defined as $\frac{1}{n}\sum CSEM_i^2$ where CSEM is the conditional standard error of the estimated θ for a student as obtained from the publisher's IRT model. The publisher reports the expected relationships among scale scores and grade levels of students, as well as

[9] In year X of the study, some schools inadvertently administered the *STAR Reading* test to their students. This led to the inclusion in our outcomes data set of a few hundred students with STAR Reading not STAR Early Literacy test scores. In these cases, we used dummy variable coding to adjust for the fact that the student had a STAR Reading score.

correlations of SEL scores to other commonly administered tests of early reading ability that are within the expected range (where the average correlation between external tests and SEL is .59).

HYPOTHESES TO BE TESTED

With SEL as the outcome of interest, we turn now to our hypotheses about the effects of BURST reading on students' early literacy learning. In what follows, we first describe our theoretical model of how the BURST program might operate to produce effects on student learning and then describe in informal terms a set of hypotheses tested with study data.

Theoretical Model. The study assigns schools to BURST treatment and then seeks to assess the effects on students' early literacy learning attributable to attendance at a treatment school. We argue that treatment schools affect students' learning as a result of provision of BURST instruction to students, where provision of BURST instruction is assumed to increase students' early literacy achievement in one of two ways. One way students experience a program effect is as result of directly receiving BURST instruction. A second way students experience a program effect is as a result of exposure to BURST instructed peers. These peer effects could occur, we argue, because students who receive BURST instruction become better learning partners to their non-instructed peers. As a result of these processes, we expect assignment of the BURST treatment to schools to lead to increases in achievement for *all* students attending BURST schools. Our analyses of implementation data discussed earlier, however, show that "struggling readers" (i.e., students who score well below or below grade-level/time of year DIBELS benchmarks) tend to receive more cycles of BURST instruction than students who score at or above DIBELS benchmark, so in addition to testing for an average effect across all students, we also test for treatment effects conditional on a student's status as a struggling reader or not. We further expect BURST effects to accumulate over time as a result of repeated exposure to treatment, where continuous enrollment at a treated school increases the odds of a student getting direct treatment and exposes all students to increased numbers of treated peers. This leads us to examine the effects of treatment on outcomes for the subgroup of pupils in our sample who were in continuous enrollment at a treatment school for three or four years. Finally, if the BURST program improves student achievement through the process of exposure to treatment, we should not expect the program to have effects on students who attend non-compliant schoools (defined here as schools that are predicted on the basis of an out-of-sample prognostic model to provide less than 1% of students with the ideal dose of 12 cycles of insruction per year). Therefore, we separately test the effect of enrollment at a treatment school on learning outcomes using only the sample of students who were enrolled at a treatment school that was expected to be compliant (as defined above).

Hypotheses. Our theoretical model leads to a set of null hypotheses to be subjected to statistical tests. These are presented in formal terms along with a full description of the specific estimation strategy to be used to test these hypotheses in Appendix A. For now, we simply state the hypotheses in null form without reference to the specific estimation or hypothesis testing approach we used. The hypotheses refer to the effects of BURST treatment—where treatment is defined as attendance at a school assigned to BURST as part of the current study.

The first set of hypotheses are intended to test for a positive and negative effect of attending a treatment school on average, that is, for *all* students. They are thus hypotheses about what are often called average treatment effects (ATE) in the research literature. The hypotheses about average treatment effects are:

- H1: The effect of BURST treatment on average = 0, to be tested against the alternative hypothesis (K1) that the average effect of BURST treatment is > 0. This a test for a positive benefit (on average) of attending a BURST school.

- H2: The effect of BURST treatment on all students = 0, to tested against the alternative hypotheses (K2) that the effect of BURST treatment is < 0. This is a test of whether or not the treatment is harmful to achievement (on average).

The next set of hypotheses refer to what are often called conditional average treatment effects (CATE) and in our study concern the effects of BURST treatment (on average) for certain subgroups of students. The first set of hypotheses concern BURST effects on "struggling readers." To simplify wording, let us define as "Red" any student whose DIBELS score falls well below DIBELS grade/time year benchmarks, "Yellow" as any student whose DIBELS score falls below (but not well below) DIBELS grade/time year benchmarks, and "Green" as any student whose DIBELS score falls at or above DIBELS grade/time year benchmarks. The hypotheses here concerning "struggling readers" are:

- H3: The effect of BURST treatment on the achievement of students who scored Red when they were first tested as part of this study by DIBELS = 0, to be tested against the alternative hypothesis (K3) that the effect for this group of students is > 0.

- H4: The effect of BURST treatment on the achievement of students who scored Yellow when they were first tested as part of this study by DIBELS = 0, to be tested against the alternative hypothesis (K4) that the effect for this group of students is > 0.

- H5: The effect of BURST treatment on the achievement of students who scored Yellow on DIBELS at *any time* in the study = 0, to be tested against the alternative hypothesis (K5) that the effect for this group of students is > 0.

The next set of hypotheses are about "Green" students. The reader will recall that BURST is intended for struggling readers and that "Green" students generally receive less BURST cycles of instruction than other students. For this reason, we want to assess the effects of BURST on this group. We also want to guard against the possibility that BURST is actually harmful for these students, as it might be if providing BURST instruction to students somehow drains instructionally-relevant resources from the regular reading program or from programs for gifted students. The hypotheses here are:

- H6: The effect of BURST treatment on the achievement of students who scored Green when they were first tested as part of this study by DIBELS = 0, to be tested against the alternative hypothesis (K6) that the effect for this group of students is > 0.

- H7: The effect of BURST treatment on the achievement of students who scored Green on DIBELS at *all* time points study = 0, to be tested against the alternative hypothesis (K7) that the effect for this group of students is > 0.

- H8: The effect of BURST treatment on the achievement of students who scored Green on DIBELS at *all* time points study = 0, to be tested against the alternative hypothesis (K8) that the effect for this group of students is < 0.

The final set of hypotheses are about exposure to treatment, once again expressed as CATE-type hypotheses. One hypothesis is about the group of students who were continuously exposed to BURST treatment, as measured by their continuous attendance for 3-4 years at a treament school, the second is about the group of students who attended schools that complied with their BURST treatment assignment (i.e., the group of students who were at schools predicted to deliver 12 cycles of instruction to at least 1% of their students, based on school characteristics during the two years prior to the first year of the study). The hypotheses here are:

- H9: The effect of BURST treatment on the achievement of students who were in continous attendance at treatment school for three or four years = 0, to be tested against the alternative hypothesis (K9) that the effect for this group of students is > 0.

- H10: The effect of BURST treatment on the achievement of students who attended a school that was predicted to comply with assignment to BURST treatment = 0, to be tested against the alternative hypothesis (K10) that the effect for this group of students is > 0.

ESTIMATION AND STATISTICAL TESTS

The hypotheses just stated are about whether or not the causal effect of BURST on the early literacy outcomes of students is greater (and in some cases less) than 0 for some defined group of students who attended treatment schools in our study. In H1 and H2, for example, we are interested in the effects of BURST for all students in treatment schools (i.e., the ATE), while in other hypotheses, we are interested in the effects of BURST on student achievement conditional on the value of some student covariate (i.e., the CATE).

Peters-Belson Approach to Estimating Treatment Effects. Researchers often estimate the effects of treatment on a focal outcome by taking the difference in mean outcomes between treatment and control groups, sometimes with adjustments for pre-treatment covariates in order to improve statistical precision. We implemented a variant of this approach arising in the work of Peters, Belson and others.[10] In this approach, we regress control group student outcomes for any observation *i* on the covariates of the student who provided that observation and then estimate the treatment effect as:

$$\hat{A}^{PB} = \mathrm{Avg}(y_{i1} - \hat{y}_{i0} \mid z_i = 1) - \mathrm{Avg}_w(y_{i0} - \hat{y}_{i0} \mid z_i = 0).$$

In this equation, z_i is an indicator of assignment to treatment rather than control, y_{i1} and y_{i0} denote student responses observed following assignment to treatment or to control conditions, respectively, and $\hat{y}_{i0}$ represents the regression model's prediction of how the same student would have responded if, potentially counter to fact, she had been assigned to control. Further, $\mathrm{Avg}(\cdot)$ denotes a simple average, whereas $\mathrm{Avg}_w(\cdot)$ denotes a weighted average.[11] When we estimate effects realized within a student subgroup (e.g., the subgroup of students whose entering DIBELS scores suggest they are a struggling reader), we additionally restrict both of

[10] For a discussion of PB estimates of causal effects, see Peters, C. C. (1941). A method of matching groups for experiment with no loss of population. *The Journal of Educational Research*, *34*(8), 606-612 and Belson, W. A. (1956). A technique for studying the effects of a television broadcast. *Applied Statistics*, 195-202. Peters's and Belson's techniques correspond to taking an average of response-minus-predicted differences over the treatment group only, without our subsequent subtraction of a weighted average calculated over the control group. In the case of our main effect estimate and several subgroup estimates, however, the control group's weighted average is 0 or nearly 0, making our estimates very close to those of Peters's and Belson's methods; thus the labeling of our estimates with "P" and "B", as in ", $\hat{A}^{PB}$."

[11] The weights used are inverse odds-of-treatment weights, which in this study differ from 1 only in the case of students entering the study through the one control school that was randomized as part of a triple along with two treatment

these averages to that subgroup. Note further that the superscript “*PB* “in $\hat{A}^{PB}$is used simply to denote that we are taking a Peters-Belson-like approach in estimation. Intuitively, we are estimating the effect of treatment on treated, where this effect is estimated as the difference between the outcomes students in the treatment group actually obtained versus what they would have been expected to obtain had they been in the control group. Appendix A provides full details of the PB-type estimates we used in our data analysis. For now, we provide a brief discussion of how these were obtained.

Our PB estimate for the average effect of BURST on students in treatment schools ($\hat{A}^{PB}$) was arrived at in five steps: (1) In the initial step, the collection of SEL outcomes observed for *control group* students in the study were regressed on a set of covariates for these students. In this step, we have one or more observations of SEL outcomes per student, and at each observation point, we predict that student’s time-specific SEL outcome from student-specific covariates such as a student’s DIBELS score at time of entry into the study, the student’s gender, age, grade, state location, and various interactions, where missing values on covariates are replaced by means after addition of dummy indicators for missingness on each covariate. The output of this regression analysis is provided in Appendix B. (2) In the next step, the coefficients from this regression model are used to “predict” the outcomes of all students, producing a $\hat{y}_{i0}$ for each observed SEL score. (3) In the next step, for each student observation, a “PB residual” is calculated as the difference between the observed outcome and the predicted outcome. (4) Separately within treatment and control groups, means of residuals are calculated. (5) The average effect of treatment on treated is estimated by subtracting this weighted mean of PB residuals for the control group from the mean of PB residuals for the treatment group.
Recall from the discussion of hypotheses to be tested during analysis that we are interested in averaging PB-type estimates across all students and across subgroups of students. When effects are being estimated for a subgroup, only residuals associated with the subgroup contribute to these means. Note also, that for hypothesis testing (only), we will be applying one of four additional weights to each observation depending on whether it is the first, second, third or fourth year of measurement of the student in question. These weights are described in the next section.

Hypothesis Testing. With PB estimates in hand, we proceeded to test the series of hypotheses listed earlier. These are one-sided tests of the null hypothesis of no effect, which we conducted as permutation tests controlling for family-wide error rates. The test statistic for the overall average effect is:

$$\frac{\sum_t \hat{w}_t \hat{A}_t^{(PB)}}{\sum_t \hat{w}_t (\#\{i \in \mathcal{F}_t : Z_i = 1\})}.$$

For all other hypotheses, simply add a subscipt to the $\hat{A}_t^{(PB)}$ term to denote the subgroup over which the average is taken. Note that these are weighted averages, which is natural given the different numbers of observations at each followup point (*t*) in the study. In a subsidiary analysis (not shown here), we found the form of these weights ($\hat{w}_t: t = 1, \ldots, 4$) that maximizes power to reject H_1 should effects at different followups be consistent with our theory of how the BURST program affects student learning. Assuming BURST effects on student learning are cumulative and proportional to students served, these optimal case weights are determined jointly by the covariance of the four time-specific effect estimates $(A_t^{\{(PB)\}}: t = 1, \ldots, 4)$ and by students’ probabilities of having been classified as Red or Yellow over the course of being observed, conditional on their assignment to the control group and on being observed over *t* years. All of this is discussed in more detail in Appendix A. Suffice it to say here that the ($\hat{w}_t: t = 1, \ldots, 4$) in the formlua above

schools; these students receive a weight of 2. For all other students this weight is 1, because all other control schools were assigned in pairs, with a priori probability $\frac{1}{2}$ of assignment to treatment.

Table 6.1 | Effects of BURST on SEL Outcomes

Panel 6.1.a | Peters-Belson Estimates of Treatment Effects (in SEL Scale Score Points) for Subgroups by Year of Follow-Up

Hypothesis	After 1 Year	After 2 Years	After 3 Years	After 4 Years	Averaged Over All Observations	Effect Size All Observations
All Students	1.07	-1.70	-.54	1.98	.03	.00
Classified Red at Start	1.17	-3.85	-3.46	-.72	-2.31	-.02
Classified as Yellow at Start	6.55	-4.44	1.34	3.37	2.11	.02
Ever Classified as Yellow	5.33	-2.17	2.03	3.92	2.37	.02
Classified as Green at Start	1.39	.36	.61	3.61	1.05	.01
Always Classified as Green	.96	1.83	-.43	4.11	1.04	.01
3-4 Years Continuous Enrollment	-1.90	.54	-.38	2.80	-.23	.00
Enrolled in Predicted Complier School	.70	-.82	1.76	2.80		

Panel 6.1.b Results of Significance Tests

	Z value	P value
All Students > 0	-.12	1.0
All Students < 0	.12	1.0
Classified Red at Start > 0	-.38	1.0
Classified as Yellow at Start > 0	.48	.95
Ever Classified as Yellow > 0	.76	.81
Classified as Green at Start > 0	.05	1.0
Classified as Green at Start < 0	-.05	1.0
Always Classified as Green > 0	-.27	1.0
Always Classified as Green < 0	.27	.99
3-4 Years Continuous Enrollment> 0	.03	1.0
Enrolled in Complier School > 0	.11	1.0

are estimates of these optimal weights calculated using sample-based estimates of the covariances and conditional probabilities just referred to.

With the test statistic in hand, reference distributions for the test statistic were tabulated by conducting 64,000 permutations of the variable recording the treatment/control distinction, then re-calculating all weights, models and differences of differences after substituting the permuted for the original treatment vector. Here the permutations take into the account the clustering of observations within students and schools and the blocking structure of random assignments. Full details of the permutation procedures, including how they controlled familywise error are provided in Appendix A.

FINDINGS ON THE EFFECTS OF BURST TREATMENT

Table 6.1a (above) presents the average PB estimates for different groups of students in treatment schools after a given number of years of followup and as averaged across all observations. In the table, these estimates are reported in SEL scale score points. An overall effect size is also presented, which is Cohen's D_{sd}. This effect size expresses the "all observations" PB estimate as a decimal fraction of the pooled, within-grade standard deviation of SEL test scores in our sample, which is 97.3.

The reader will immediately see that none of the estimated effects of BURST on students' early literacy achievement is large. Take, for example, the data reported in the first row of Table 6.1a, which reports PB estimates at different followup points for *all* students in the treatment group. At the end of one year in the study, students scored on average 1.07 SEL scale score points higher than would be predicted had they been in the control group, and as years of followup increase, the PB estimates vary around zero such that, over all years, the average PB estimate is just .03 scale score points or a D_{sd} of essentially zero. The results are not much different for any of the subgroups either. The highest PB estimates appear in the first year of the study for Yellow students—but these are small in both absolute terms (about 5 or 6 SEL scale score points) and not confidently different from zero. In fact, no effects are large (in either a positive or negative direction), and when averaged across all years of the study, the effects on all or any group of students of having been in a BURST treatment school are never larger in terms of Cohen's D_{sd} than .02. Given the size of these effects, the significance test results shown in Table 6.1b are expected. Not a single test confirmed that BURST treatment effects were greater than (or less than) 0 for any group of pupils on whom a hypothesis was tested.

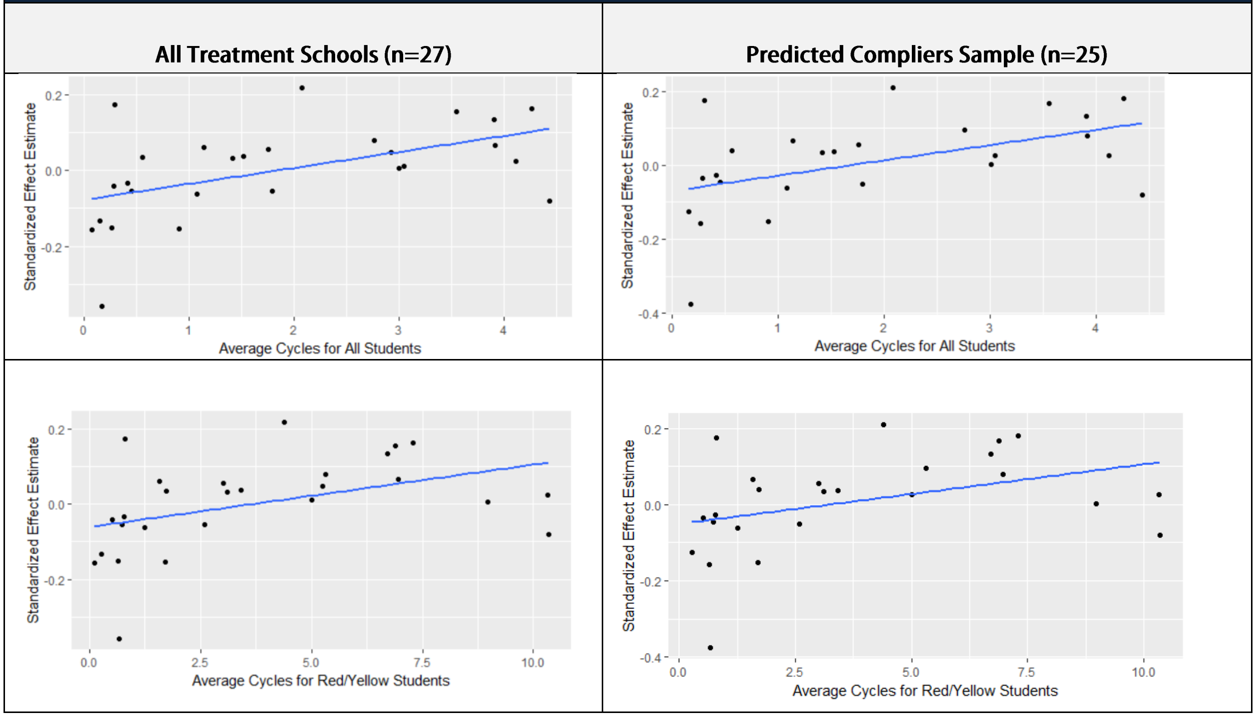
AN EXPLORATORY ANALYSIS

Our theory of the BURST program and how it might affect students' early literacy learning implied that BURST effects should be larger in schools that provide more BURST instruction to students. As a result, we conducted an exploratory analysis of this issue. The analysis is exploratory because schools in this study were *not* randomly assigned to provide more or fewer cycles of BURST instruction, and so any correlation that arises between a school's allocation of more or fewer BURST cycles and PB estimates of program effects could be due to some omitted variable that acts as a "common cause" of both student achievement and the amount of BURST instruction offered to students. In particular, we can imagine the possibility that some schools have more capacity than others to mount high quality instruction and that this unobserved capacity affects both how many cycles of BURST instruction they offer to schools and PB effect estimates. With that caution firmly in mind, we now report on the correlation between school-level PB estimates of BURST effects and the average number of cycles of BURST instruction offered to students in a school.

Figure 6.2 (next page) presents four scatterplots of these data. In the left hand scatterplots, all treatment schools are in the data; in the right hand scatterplots, the 2 schools predicted to be non-compliers are omitted from the data. In all graphs, the X (or horizontal) axis is the average number of cycles of BURST instruction offered to students per year, while the Y (or vertical) axis re-scales the PB estimates as standardized effect sizes (as was done in Table 6.1a). The reference line in each scatterplot is the least squares regression line. All of the scatterplots show a positive correlation between average BURST cycles provided to students and PB effect estimates. In the "all schools" samples on the left, the rank order correlation (Kendall's τ) between these variables is around .40 for all students. On the right, it is .38 for Red/Yellow students. In the "compliers only" sample on the right hand side, the rank order correlation is around .35 for both scatterplots. There is thus a small, positive relationship between average cycles of BURST instruction offered and school average PB estimates. But none of the scatterplots suggest that within the experimental sample, an increase in the average number of cycles of BURST instruction offered at a school was accompanied by substantial increases in program effects. Indeed, from the reference line, it can be seen that at the lowest levels of implementation, the expected PB estimates are just below zero, while at the highest levels of implementation, the expected

PB estimates are around .10 (implying a standardized effect size at this point on the regression line of around D_{sd} = .10). Bear in mind that the scatterplots are presented without confidence intervals on estimates, so we cannot confidently say that effects observed at any point on the reference line are confidently different from zero. We can, however, conclude that expected PB effects are quite small at all points on the regression line, implying that schools that offer more BURST instruction on average are not expected to increase students' early literacy learning by much more than would be observed in schools that used only DIBELS as a formative assessment within their instructional programs.

Figure 6.2 | Scatterplots of Relationship Between Average # of Cycles of BURST Instruction Offered at a Treatment School and School PB Estimates



All Treatment Schools (n=27)

Predicted Compliers Sample (n=25)

7 | DISCUSSION OF FINDINGS

The analyses presented in the last section strongly suggest that *BURST®: Reading* had no discernable positive (or negative) effects on the early literacy achievement of students attending schools assigned to the BURST treatment as part of this study. Indeed, across all four years of the study, the effects on all students who attended a BURST treatment school, the effects for "struggling readers" who attended a BURST treatment school, the effects for students who were at or above grade level in early reading skills and who attended a BURST school, the effects for students who attended BURST schools for three to four years continuously, and the effects for students who attended schools that were predicted to comply with their assignment to BURST treatment were *never* confidently greater (or, in cases where an additional statistical test was conducted, confidently less) than zero for any group of students considered. We now discuss the larger inferences we think readers can make from these findings about the effectiveness of the *BURST®: Reading* program.

Consider the Counterfactual When Reporting Study Results

The effects just described are based on a statistical model comparing the early reading of achievement of students at BURST schools to what that statistical model predicted those same students would have achieved had they been in a control school in the study. Importantly, in the current study, the control condition was not a "business as usual" control group but rather a control group of schools in which students were tested at similarly high rates with DIBELS as in the treatment group. When readers discuss the effectiveness of the BURST program, they should therefore keep the nature of the control group firmly in mind. In the current study, we found that the added benefit of using BURST *over and above a process of universal screening with DIBELS* was negligible.

Consider How BURST Was Implemented When Reporting Study Results

The reader will also recall that, in all treatment schools in this study, BURST instruction was offered to students at rates that were below what the vendor (Amplify) considers ideal. As discussed in Sections 2 and 5 of this report, Amplify recommends that schools offer struggling readers 12 cycles of BURST instruction a year. In this study, however, the average student who received BURST instruction in a given year was exposed to just 6 cycles of BURST instruction in that year. Moreover, not all struggling readers received BURST services. In this light, the BURST Efficacy trial reported here should not be considered a test of how effective BURST is under "ideal" conditions of implementation. Rather, it is a test of how effective BURST instruction is when implemented under more "routine" conditions of implementation such as would be observed in schools that purchase BURST in a market transaction with Amplify and use it according to local capacity and circumstances. Indeed, in Section 5 of this report, we offered evidence of how average patterns of BURST implementation in treatment schools resembled average patterns of use observed in similar schools that had purchased BURST outside the current study and were using it under routine conditions of implementation in AY2016-2017. The point to be taken from that analysis is that when readers discuss the effectiveness of the BURST program, they should keep firmly in mind that this study found the added benefit of using BURST over and above a process of universal screening with DIBELS was negligible *under "routine" conditions of implementation*.

The findings on levels of BURST service provision in treatment schools warrant further discussion. Amplify advises schools to offer BURST instruction to all struggling readers, and over the course of the year, Amplify suggests that struggling readers get 6 cycles of BURST instruction a semester or 12 cycles of instruction per year. This advised pattern of BURST implementation is certainly feasible within the normal school calendar and can be achieved by offering identified students supplementary reading instruction 2-3 times per week (even allowing for weeks when no supplementary instruction is offered). But, as Figure 6.2 showed, only two schools in the treatment group offered this level of BURST instruction to their struggling readers, while most schools offered less. A problem that needs to be taken up, then, to explain why there was a pattern of "less than ideal" patterns of service provision in treatment schools.

One plausible explanation—and the one we elaborate on here—is that patterns of BURST service provision observed in study schools resulted from the resource constraints schools faced. On this view, study schools had more struggling readers than they had the capacity to serve at levels of service provision recommended by Amplify. We know from the data reported in Table 5.4, for example, that in the *peak* year of service delivery (year one of this study), the average school managed to provide BURST instruction to a little more than 50% of struggling readers in a given semester and we know from other data, that many students served in the first semester were not served in the second semester of a given year. This looks to us, then, like a pattern of "rationing" in which study schools could allocate *all* students in need of services to functioning BURST groups for lack of resources and that, as a result, study schools attempted to strike a balance between serving fewer students with more cycles versus serving more students with fewer cycles. The potential problem of "rationing," then, is all the more reason to report this study as having found that *under "routine" conditions of implementation* the added benefit of using BURST over and above a process of universal screening with DIBELS was negligible. Moreover, the reader should understand that this statement draws *no* conclusions about the effects of using BURST under "ideal" conditions of implementation. The current study did not conduct a test for the effects of BURST under ideal conditions.

Consider the Kinds of Schools in the Treatment Sample When Reporting Study Results

The idea that routine conditions of implementation involve constraints on BURST provision raises another issue that should be taken into account when reporting the results of this study. As discussed in Section 3 of this report, the collection of schools in the BURST treatment group was not representative of all schools in the U.S., nor was it representative of all schools that had purchased and were using BURST in AY 2016-2017 (i.e., the BURST "user population"). On average, schools in the treatment group were located in communities that were poorer than both the average U.S. community and the communities in which BURST user group schools were located. In addition, schools in the treatment group were located in smaller school districts than both the average U.S school district and the average district in which BURST user group schools were located. Finally, students in treatment group schools were lower achieving than students in both the average U.S. school and schools in the BURST user group. All of this could account (in part) for the pattern of rationing discussed above, as well as for why, in comparison to schools in the BURST user group, treatment grou8p schools in this study were at the *higher* end of percentage of students receiving any BURST instruction but at the *lower* end of percent of students receiving a "full" dose of BURST reading cycles. Being located in poor communities with high percentages of poor, lower achieving students, treatment schools seemingly "rationed" the allocation of BURST instruction to struggling readers. For this reason, it makes sense to view the results of this study as suggesting that *in schools located in small school districts, in poor communities, with high percentages of lower achieving students*, the added benefit of using BURST (over and above a process of universal screening with DIBELS) was negligible under routine conditions of implementation.

Implications of Study Results for School Improvement

Finally, let us consider the implications of this study for the larger question of how to improve the early literacy learning of "struggling readers"—especially struggling readers living in poor communities served by smaller school districts in which schools have many struggling readers and are using DIBELS as a universal screening tool. In this situation, schools have many choices for the kinds of curricula they might purchase for use in their supplementary reading programs and in how they use that curricula with different groups of pupils. The data presented here suggest that *the use of the BURST curriculum and student grouping algorithm with struggling readers is likely to be no better or worse a choice than the use of alternative curricula and grouping practices that schools might use* and, thus, that a decision about whether or not to purchase and use BURST under routine conditions of implementation can, without much consequence for student achievement, be made on the basis of instructional preferences and cost considerations.

Appendix A

IES study of BURST: Student achievement outcome analysis plan

Ben Hansen

18 October, 2018

Overview

The primary research questions of the IES funded RCT studying the Burst early literacy support program call for comparisons of treatment and control scores on the Star Early Literacy examination, administered as a post-test at study schools in each of grades K-3 and in each of the 4 years of the study. This note details our plans for analysis of this student outcome, with special attention to novel aspects of our analysis plan and methods.

Our analyses of this outcome each adjust for several pre-treatment covariates. The covariates for which we make adjustments are commonly considered in education research, but our mode of adjustment differs from common approaches in fitting its covariance model separately from and prior to the comparative analysis of outcomes, an approach originating with Peters (1941), Belson (1956) and Cochran (1969). Its separation of covariance and comparative outcome analysis facilitates its estimation of the benefit (or harm) that may be attributed (Paul R. Rosenbaum 2001) to the Burst progam. A novel approach to combining program effect estimates at different follow-up points is employed to enhance the power of the main hypothesis test aiming to detect a program benefit, if in fact the program is beneficial. Power is estimated through a linked simulation study, conducted following data collection but prior to estimation of treatment effects. The results of this simulation study also informed the specification of the covariance model.

One and the same covariance model assists inferences about treatment effects overall and separated out by time of observation, or by student subgroup; these randomization-based inferences are arranged within a comprehensive multiple-comparisons strategy centering around the Romano-Wolf (2005) step-down procedure and ensuring family-wise error rate control for a family of tests for overall or by-subgroup program benefits, as well as over subsequent tests for differences in program efficacy by subgroup. These "confirmatory" moderation analyses are supplemented by a range of additional tests of additional moderation- and mediation-related hypotheses. Although "exploratory" rather than confirmatory, these tests are enumerated and pre-declared in the study analysis plan, enabling us to appraise significance using the Benjamini-Hochberg (1995) step-down p-value adjustment, and thus accompany any findings in this domain with an estimate of a corresponding false discovery rate (Benjamini and Yekutieli 2001; Efron 2012).

Attributable effects

Covariance adjustment in the Peters-Belson-Cochran mode

Our covariance model is fit to observations on control group students using an ordinary multiple regression, pooling all available observations on the control group within the same model (regardless of student subgroup or time of observation). Here "control group" refers to students joining the study population by way of enrollment at a school assigned to control – that is, on students whose "join school," the study school at which they were first observed, had been randomly assigned to the control condition. (Irrespective of whether the student remained at that school for the duration of the study.) Subsequent to the fitting of this model, model predictions $\hat{\mu}_{Ct}$ are obtained for both treatment and control group students, one for each time point at which a student was observed.

Estimation of attributable effects

Within the conceptual framework of potential outcomes (Neyman 1923; Rubin 1974; Holland 1986), any outcome observation Y_i on a student who happens to have been assigned to treatment, $Z_i = 1$, may be separated into what would have been observed in the absence of the treatment, y_{Ci}, plus a component attributable to the treatment, $y_{Ti} - y_{Ci}$. (As an immediate consequence of the identity $Y = ZY_T + (1-Z)Y_C$.) Of course neither component is observed in isolation from the other. However, within a collection of observations $\mathcal{F}$ that is defined without reference to treatment assignment, the total of individual outcome components attributable to the treatment,

$$A_{\mathcal{F}} = \sum_{i \in \mathcal{F}} Y_i - y_{Ci} = \sum_{i \in \mathcal{F}: Z_i = 1} y_{Ti} - y_{Ci},$$

known as an attributable effect (Paul R. Rosenbaum 2001), can be an object of statistical inference in the conventional frequentist sense. (This despite its status as a random variable rather than a fixed parameter.) For example, writing o_i for the treatment assignment odds associated with observation i[1], the observable quantity

(1) $\sum_{i \in \mathcal{F}: Z_i = 1} y_{Ti} - \sum_{i \in \mathcal{F}: Z_i = 0} o_i y_{Ci} = \sum_{i \in \mathcal{F}: Z_i = 1} Y_i - \sum_{i \in \mathcal{F}: Z_i = 0} o_i Y_i$

estimates $A_{\mathcal{F}}$ without bias, and in common designs the sampling variance of its difference with $A_{\mathcal{F}}$ can also be estimated without bias. Estimator (1) is shown here for illustrative purposes only; the study will instead use the Peters-Belson type estimator

$$\hat{A}^{(PB)} = \sum_{i \in \mathcal{F}: Z_i = 1} Y_i - \hat{\mu}_{Ci} - \sum_{i \in \mathcal{F}: Z_i = 0} o_i (Y_i - \hat{\mu}_{Ci})$$

where $(\hat{\mu}_{Ci} : i)$ represent predictions of the outcome that emerge from using student observations on the control group to estimate a model of $\mathrm{E}(Y_C | \mathbf{X}, Z = 0)$, then combining the fitted model with $\mathbf{x}$-values of both treatment and control observations to generate regression predictions

$$\hat{\mu}_{Ci} = \hat{\mathrm{E}}(Y_C | \mathbf{X} = \mathbf{x}_i).$$

But for the following differences, this is the estimation technique discussed by B. B. Hansen and Bowers (2009):

- Hansen and Bowers used logistic regression to model $\mathrm{E}(Y_C | \mathbf{X}, Z = 0)$; the Burst analysis will use ordinary least squares;
- When standard errors are needed, the Burst analysis uses a different variance estimation procedure (describe in a footnote below).

When $\mathcal{F}$ is the collection of all available follow-up observations, $A_{\mathcal{F}}$ might be referred to as *the* attributable effect, rather than *an* attributable effect; but because we will often want to separate observations according to year of follow-up, if not also more finely, we retain the restriction to a generic collection of observations $\mathcal{F}$.

In calculating t-statistics used to test whether program effects differ by subgroup, or in additional moderation analyses, these attributable effect estimates are scaled by standard errors determined using a variant[2] of the method of B. B. Hansen and Bowers (2009). Estimates of average treatment effects in the intention to treat

[1] That is, the *a priori* odds of assignment to treatment, as opposed to control, governing the randomization process indirectly determining the treatment condition under which observation i was obtained. The relevant probabilities being a function of the randomization block B_i containing the join school of the student on whom (x_i, y_i) is an observation, $\Pr(Z_i = 1 | B_i) / \Pr(Z_i = 0 | B_i)$.

[2] In both cases the standard error is calculated with attention both to random assigment blocks and to clustering of observations within blocks, and in both cases the standard error calculation makes use of outcome data for the control group but not for the treatment group. In both cases the variance an overall attributable effect is represented as a sum of variances of block-specific attributable effect contributions, with the overall attributable effect variance estimated by summing separate estimates of blockwise variance contributions. However, in Burst but not in the studies discussed by B. B. Hansen and Bowers (2009), in all but one of the blocks the control condition is represented by a single cluster, rendering inapplicable the blockwise variance estimators used by B. B. Hansen and Bowers (2009), as well as many others (Pashley and Miratrix 2017). In light of this, for the Burst study blockwise estimates of variance are fashioned from a scaling of the "bread" component of Huber-White/sandwich type estimate of the variance associating with a weighted mean of student-wise Peters-Belson outcome residuals, where:

sense, overall or by subgroup, are simply ratios of estimated attributable effects to numbers of treatment group observations falling within the subgroup. Under the additional assumption of an exclusion restriction, complier average treatment effects are estimated as ratios of attributable effects to the number of compliers observed in the corresponding treatment groups.

Dry-run simulation analysis

To inform the specification of the Peters-Belson covariance adjustment model, and to confirm that the Peters-Belson/attributable effects strategy need not sacrifice power relative to estimation strategies more commonly employed in education research, we explored statistical performance of these methods within simulated experimental comparisons. The simulations featured naturalistic data generating mechanisms built in part from the same data resources that the actual experiment requires for the fitting of a Peters-Belson covariance model, namely school and student covariates along with student outcomes, but only for control group students. (Conducting these simulation studies did not require us to break our "self-blind" on treatment group outcomes.) The design of the simulation experiment, adapting to the repeated-measures cluster randomized trial setting the "dry-run analysis" strategy described in Wyss et al. (2017), will be described in detail elsewhere.

Summative null-hypothesis test; weighting scheme employed for this purpose

Let $\mathcal{F}_t$, $t = 1, 2, 3, 4$, refer to the collection of observations obtained after t years of follow-up. (Thus $\mathcal{F}_1$ contains all observations from study year 1 along with additional observations from each of study years 2, 3 or 4, as new students joined the study, whereas $\mathcal{F}_4$ contains only observations from year 4, and then only on the subset of year 4 participants who were observed and tested in each of years 1-4.) Abbreviate $A_{\mathcal{F}_t}$ by A_t, $t = 1, \ldots, 4$. For the purpose of testing the no-effect hypothesis that *each* of

$$\mathrm{E}(A_t) = \mathrm{E}(Y_T - Y_C|\mathcal{F}_t, Z = 1) \cdot \Pr(Z = 1|\mathcal{F}_t) \cdot (\#\mathcal{F}_t), \; t = 1, \ldots, 4,$$

is *zero or negative*, against the alternative hypothesis that *one or more* of these scaled treatment effect averages is *positive*, it is natural to select weights $w_1, w_2, w_3, w_4 \geq 0$, rejecting the no-effect hypothesis for sufficiently large values of $\sum_{t=1}^{4} w_t A_t$.

It is also natural to select these weights in such a manner to maximize the power of the resulting test, in the event that the null hypothesis is true.
Statistics of form $\sum_{t=1}^{4} w_t A_t$ being approximately Normal, for any simple alternative K power is maximized by taking $(w_t : t)$ to maximize the quantity

$$(2) \quad \frac{\mathrm{E}_K\left(\sum_{t=1}^{4} w_t A_t\right)}{\left\{\mathrm{Var}_K\left(\sum_{t=1}^{4} w_t A_t\right)\right\}^{1/2}}.$$

For all K according to which treatment effects are positive and

$$\mathrm{E}(Y_T - Y_C|\mathcal{F}_t, Z = 1) \propto p_{Ct}, \; t = 1, \ldots, 4,$$

where

$$p_{Ct} := \Pr(\text{if assigned to control, student would `test in' by time}\, t|\mathcal{F}_t),$$

it turns out that (2) is maximized by the same weights $(w_t : t)$. (Students "test in" to the Burst intervention by scoring at or below grade and time of year specific benchmarks on the DIBELS:NEXT examination,

- the mean is taken over the control condition only; - student-year observation i is weighted by o_i, the corresponding odds of assignment to the treatment condition; - observations are clustered on the randomization block, B_i, associated with the student in question. (Whenever the student is associated to a school that was randomly assigned as part of a pair, or as part of a triple assigning only one school to control, "clustering on randomization block" is the same as clustering by school; but for students joining the study through a school belonging to a randomization block with three schools, two of which were assigned to control, a single psuedo-cluster is formed from what in actuality are two distinct clusters.) This cluster-aware covariance estimate will be effected using the method of Pustejovsky and Tipton (2017).

administered electronically at study schools over the course of the study; these events are recorded in study data.) The optimal weights take the form

$$(\text{Const}) \cdot \Sigma^{-1}\mathbf{p}_C, \text{ where } \Sigma := \text{Cov}\left\{(\hat{A}_t^{(PB)} - A_t : t)\right\},$$

(Const) being a positive constant, $\mathbf{p}'_C$ being the column vector of proportions p_{Ct}, $t = 1, \ldots, 4$, and $\hat{A}_t^{(PB)}$ being the result of our Peters-Belson procedure as applied to estimation of attributable effects specific to follow-up time t. Although neither $\mathbf{p}_C$ nor Σ is directly observed, both are estimable, each p_{Ct} as a proportion $\hat{p}_t$ observed among students assigned to control and Σ by a somewhat more elaborate calculation[3] that also calls only for data about the control group. Thus we can calculate $\hat{\Sigma}$ and $\hat{\mathbf{p}}_C$ before removing our self-blind of information about outcomes within the treatment group, yielding (estimated) optimal case weights $\hat{\mathbf{w}} = \hat{\Sigma}^{-1}\hat{\mathbf{p}}_C$.

A leading component of our assessment of evidence for a benefit from the intervention will be a one-sided test of the hypothesis of no effect. This test will be a permutation test, using as its test statistic

$$\frac{\sum_t \hat{w}_t \hat{A}_t^{(PB)}}{\sum_t \hat{w}_t (\#\{i \in \mathcal{F}_t : Z_i = 1\})}.$$

For the purpose of tabulating a reference distribution of this test statistic, $(\hat{w}_t : t)$ and $(\hat{A}_t^{(PB)} : t)$ are recalculated following each permutation of the observations. Students' treatment conditions are re-randomized indirectly, by reassigning treatment allocation labels to schools rather than students, in this fashion respecting the study's clustering of observations within student and school. At least 1,000 random permutations of school-condition associations will be made. Replicate assignments of schools to conditions will follow the treatment assignment probabilities and blocking structure that in actuality was used to make random assignments, furnishing a finite-sample correct, randomization-based p-value for the hypothesis of strictly no effect.

In order to integrate this test with by-subgroup tests for presence of treatment effects, described below, a permutation p-value for the hypothesis of an overall benefit of treatment will be determined by the following procedure, which is engineered to reject at level 0.05 whever a one-sided test of an overall benefit would have given a rejection at level 0.025, while also to control type 1 error at 0.05 over a broader family of hypotheses (to be described in subsequent sections):

- The permutation covariance of this test statistic and subgroup-specific statistics, described below, will be approximated by rerandomizing and calculating the statistics' covariance across random assignments.
- Scaling each of these statistics by its approximated permutation s.d. furnishes a vector of z-statistics, and corresponding approximate correlation matrix. Let z_1 be the test statistic described above, scaled by its permutation s.d.. Determine the upper 0.025 quantile, i.e. the 97.5th percentile, of its null permutation distribution: we expect this number to be 1.96, based on the Normal approximation; if it's something different, replace "1.96" in bullet points below with that other number.

[3] In both cases the standard error is calculated with attention both to random assigment blocks and to clustering of observations within blocks, and in both cases the standard error calculation makes use of outcome data for the control group but not for the treatment group. In both cases the variance an overall attributable effect is represented as a sum of variances of block-specific attributable effect contributions, with the overall attributable effect variance estimated by summing separate estimates of blockwise variance contributions. However, in Burst but not in the studies discussed by B. B. Hansen and Bowers (2009), in all but one of the blocks the control condition is represented by a single cluster, rendering inapplicable the blockwise variance estimators used by B. B. Hansen and Bowers (2009), as well as many others (Pashley and Miratrix 2017). In light of this, for the Burst study blockwise estimates of variance are fashioned from a scaling of the "bread" component of Huber-White/sandwich type estimate of the variance associating with a weighted mean of student-wise Peters-Belson outcome residuals, where:
- the mean is taken over the control condition only; - student-year observation i is weighted by o_i, the corresponding odds of assignment to the treatment condition; - observations are clustered on the randomization block, B_i, associated with the student in question. (Whenever the student is associated to a school that was randomly assigned as part of a pair, or as part of a triple assigning only one school to control, "clustering on randomization block" is the same as clustering by school; but for students joining the study through a school belonging to a randomization block with three schools, two of which were assigned to control, a single psuedo-cluster is formed from what in actuality are two distinct clusters.) This cluster-aware covariance estimate will be effected using the method of Pustejovsky and Tipton (2017).

- Assuming a multivariate Normal random vector Z statistics with mean zero and covariance equal to that correlation matrix, find $c > 1$ such that the 95th percentile of $\max(c \cdot Z_1, \max_{k>1} Z_k)$ equals $1.96c$. (As $c \uparrow \infty$ the 95th percentile of $\max(c \cdot Z_1, \max_{k>1} Z_k)$ approaches 1.64, so this is possible.)
- The p-value associating with the the hypothesis of strictly no effect is that determined by application of the step-down max-T procedure of Romano and Wolf (2005). However, the calibration above ensures that it is rejected at level .05 if $z_1 > 1.96$, just as a one-sided level .025 test using a normal approximation (but without multiplicity corrections) would reject if $z_1 > 1.96$.

Caveats about power as reported in our SREE Registry entry

1. Being based on our dry-run simulation study, power calculations we report in our SREE Registry entry are informed by covariate and outcome data from the control group, but no outcomes from the treatment group.

2. The power and minimum detectable effect size (MDES) approximations reported there summarize results of simulating treatment-group outcomes under alternative hypotheses granting treatment effects only to students who have tested in. From each simulated data set, the hypothesis is tested not with a permutation test but by referring $\left(\hat{\mathbf{w}}\hat{\Sigma}\mathbf{w}'\right)^{-1/2} \sum_t \hat{w}_t \hat{A}_t^{(PB)}$ to a Student's t-distribution — an approximation to the permutational procedure planned for the actual study.
3. The power and MDES approximations we report there are based on testing a one-sided null at level 0.025, one "half" of a conventional, symmetric two-sized test at level 0.05. In actuality we plan to conduct a two-sided test, and for this to be one "half"" of it. I.e., for this two be one of two tests, of the same null but against different alternatives, with p-values corrected for multiplicity according to the Bonferroni principle. However, our two-sided test will be asymmetric, with its other half itself being a test of a composite of hypotheses, the refected image of the primary test (for the presence of a treatment benefit) being one just one among several combined in this family. These additional hypotheses to be tested as a part of the same family will be described in the next section. Because of the multiplicity of tests within the family as a whole, the resulting two-sided test of $H : \mathrm{E}(A_t) = 0, t = 1, \ldots, 4$ is asymmetric in the sense that test statistics $\sum_t \hat{w}_t \hat{A}_t^{(PB)} = \pm t_0$, $t_0 > 0$, would not occasion the same two-sided p-value, the p-value corresponding to a $\sum_t \hat{w}_t \hat{A}_t^{(PB)} = -t_0$ being larger.
4. Were it our plan to test the hypothesis of no effect with a symmetric two-sided test, the power calculation procedure used for our SREE registry entry would underestimate power against positive-effect alternatives by a small, effectively negligible amount. But that is not our plan: although we will test against alternatives asserting a negative effect, we will not do this in a symmetric fashion, allocating half of our type 1 error budget to these tests. Rather, we'll include the test for a negative main effect among an array of subgroup-specific tests for the presence of subgroup-specific effects to be described in the next section. Because most of the subgroup-specific tests within this battery aim to detect positive effects, and might in principle do so even if the "main effect" test discussed within this section does not, there is an additional contribution to our procedure's aggregated power to detect a benefit of the program. This addition is of indistinct magnitude.

Tests for subgroup-specific treatment effects and effect moderation, and of hypotheses relating to mediation

As noted immediately above, in parallel with testing for the presence of a (positive) treatment effect as averaged across all follow-up observation times and subgroups, we will also test for treatment effects within several student subgroups. Considered as Neyman-type tests with a fixed accept/reject criterion, and conventional Type 1 error allowance 5%, these tests are coordinated in such a way that they and the "main effect" test described above may be considered as a test family, with corresponding family-wise error rate (FWER) 5%. In the language of NCEE guidelines pertaining to multiple testing (Schochet 2008), these are *confirmatory* tests.

Subsequent but related procedures consider for each of the same student subgroups whether effects of the treatment differ between the subgroup and its complement. For multiple-comparisons purposes these tests also are considered to be confirmatory, and to belong to a common family for which the FWER is to be controlled.

We will then go on test a range of additional hypotheses relating to moderation and mediation. Although we consider these tests to be exploratory, in the NCEE guidelines' (Schochet 2008) sense, we are pre-specifying them with this registry entry, and p-values that we report for them will have been adjusted for false discovery rate control.

The following notation is used in this section.

- For each follow-up time $t = 1, \ldots, 4$, let $\bar{w}_t = \mathrm{E}(\hat{w}_t)$, where $\hat{w}_t$ the estimated power-optimizing weight described above and expectation is evaluated relative to the rerandomization distribution.

- Let $\mathcal{G}^{(start)}$, $\mathcal{Y}^{(start)}$ and $\mathcal{R}^{(start)}$ be the collections of all observations on students whose earliest available DIBELS:NEXT scores were, respectively, at or above the relevant grade and time of year-specific benchmark (i.e. "green"), "below benchmark" (yellow), or "well below benchmark" (red).

- Let $\mathcal{Y}^{(ever)}$ consist of observations on the subset of study participants who at some point during the course of the study test below benchmark on DIBELS:NEXT, and whose first such score was "below benchmark" or yellow rather than "well below benchmark" (red).
- Let $\mathcal{G}^{(always)}$ indicate observations on the subset of study participants who are only observed to score at or above benchmark on DIBELS:NEXT score, *and* are observed for 3 or 4 years at the same school.
- Let $\mathcal{C}$ consist of all observations on study participants who over the course of the study are observed to continously attend the same study school for 3 or 4 consecutive years, and at some point during that time test red or yellow on DIBELS:NEXT, i.e. below benchmark.
- Let $\mathcal{G}_1^{(start)} = \mathcal{G}^{(start)} \cap \mathcal{F}_1$, the subset of student observations in $\mathcal{G}^{(start)}$ that are at the first year of follow-up for that student, and for $t = 2, 3, 4$ let $\mathcal{G}_t^{(start)} = \mathcal{G}^{(start)} \cap \mathcal{F}_t$ be the subset of observations within $\mathcal{G}^{(start)}$ that are at the t year of follow up for the student in question. Similarly define $\mathcal{Y}_t^{(start)}$, $t = 1, \ldots 4$, and so on through $\mathcal{C}_t$, $t = 1, \ldots, 4$.

A family of hypothesis tests regarding whether and how the program is beneficial

The following null hypotheses will be tested as a family, using a max-T type procedure (Hothorn, Bretz, and Westfall 2008) to generate p-values adjusted to control FWER.
0. H: $\sum_t \bar{w}_t \mathrm{E}(A_t) = 0$, against the alternative K: $\sum_t \bar{w}_t \mathrm{E}(A_t) > 0$. (This is the test for the presence of a benefit described above.)

1. H: $\sum_t \bar{w}_t \mathrm{E}(A_t) = 0$, against the alternative K: $\sum_t \bar{w}_t \mathrm{E}(A_t) < 0$, i.e. that treatment is harmful on average.
2. H: $\sum_t \bar{w}_t \mathrm{E}(A_{\mathcal{G}_t^{(start)}}) = 0$, against alternative K: $\sum_t \bar{w}_t \mathrm{E}(A_{\mathcal{G}_t^{(start)}}) > 0$.
3. H: $\sum_t \bar{w}_t \mathrm{E}(A_{\mathcal{G}_t^{(start)}}) = 0$, against alternative K: $\sum_t \bar{w}_t \mathrm{E}(A_{\mathcal{G}_t^{(start)}}) < 0$.
4. H: $\sum_t \bar{w}_t \mathrm{E}(A_{\mathcal{Y}_t^{(start)}}) = 0$, against alternative K: $\sum_t \bar{w}_t \mathrm{E}(A_{\mathcal{Y}_t^{(start)}}) > 0$.
5. H: $\sum_t \bar{w}_t \mathrm{E}(A_{\mathcal{R}_t^{(start)}}) = 0$, against alternative K: $\sum_t \bar{w}_t \mathrm{E}(A_{\mathcal{R}_t^{(start)}}) > 0$.
6. H: $\sum_t \bar{w}_t \mathrm{E}(A_{\mathcal{Y}_t^{(ever)}}) = 0$, against alternative K: $\sum_t \bar{w}_t \mathrm{E}(A_{\mathcal{Y}_t^{(ever)}}) > 0$.
7. H: $\sum_t \bar{w}_t \mathrm{E}(A_{\mathcal{G}_t^{(always)}}) = 0$, against alternative K: $\sum_t \bar{w}_t \mathrm{E}(A_{\mathcal{G}_t^{(always)}}) > 0$.
8. H: $\sum_t \bar{w}_t \mathrm{E}(A_{\mathcal{G}_t^{(always)}}) = 0$, against alternative K: $\sum_t \bar{w}_t \mathrm{E}(A_{\mathcal{G}_t^{(always)}}) < 0$.
9. H: $\sum_t \bar{w}_t \mathrm{E}(A_{\mathcal{C}_t}) = 0$, against alternative K: $\sum_t \bar{w}_t \mathrm{E}(A_{\mathcal{C}_t}) > 0$.

Each hypothesis will be tested using a permutation test, scaling

$$\frac{\sum_t \hat{w}_t \hat{A}_{\mathcal{D}_t}^{(PB)}}{\sum_t \hat{w}_t (\#\{i \in \mathcal{D}_t : Z_i = 1\})},$$

where $\mathcal{D} = \mathcal{G}^{(start)}, \ldots,$ or $\mathcal{C}$ as appropriate, by a (resampling-based estimate of) its permutation standard deviation under the null hypothesis to obtain a z-statistic. Considering any one of these tests in isolation, its z-statistic could be converted into a local p-value; considering them as a family, they play the role of a collection of t-statistics in a max-T calculation (Romano and Wolf 2005; Hothorn, Bretz, and Westfall 2008). This max-T procedure combining the local tests generates p-values $p_1^{(1)}, p_2^{(1)}, \ldots, p_9^{(1)}$, together maintaining FWER control with respect to the hypothesis family $\{H_1, \ldots, H_9\}$: up to error due to the use of large-sample approximation, $\Pr(\exists i \leq 9 : H_i \text{ is true}, p_i^{(1)} \leq p_0) \leq p_0$, any $p_0 \in (0, 1)$. The adjusted p-values $p_1^{(1)}, p_2^{(1)}, \ldots, p_9^{(1)}$ will be larger than the "local" p-values $p_{(1)}, \ldots, p_{(9)}$ than would have been obtained from ordinary Student's t calculations that do not situate each test within a family.

For the purpose of determining the p-value on the intersection of H_0, H_1, $\ldots$, H_9 — H_0 denoting the comparison of null $\sum_t \bar{w}_t \operatorname{E}(A_t) = 0$ to alternative K: $\sum_t \bar{w}_t \operatorname{E}(A_t) > 0$, i.e. the test for an overall benefit — $p_1^{(1)}, p_2^{(1)}, \ldots, p_9^{(1)}$ are futher adjusted, combined with the one-sided p-value p_0 from the test of H_0 described above according to the Bonferroni principle. I.e., they're doubled, giving

$$p_0^{(2)} = 2p_0,\ p_1^{(2)} = 2p_1^{(1)},\ p_2^{(2)} = 2p_2^{(1)},\ \ldots, p_9^{(2)} = 2p_9^{(1)}$$

and asymmetric two-sided p-value $\max_{0 \leq i \leq 9} p_i^{(2)}$.

A second family of hypothesis tests regarding whether program effects differ by subgroup

For each of 2–9 above we also plan a test of the hypothesis that the treatment effect for the subgroup in question is the same as the treatment effect for the complement of that subgroup. These are necessarily Wald tests rather than permutation tests; we'll effect them using bootstrap-t calculations. The direction of each of these hypothesis tests parallels that listed above. For example, in parallel with H_2 above H_2' will refer to a test of the null hypothesis that

$$\left\{\frac{\sum_t \hat{w}_t A_{\mathcal{G}_t^{(start)}}}{\sum_t \hat{w}_t(\#\{i \in \mathcal{G}_t^{(start)} : z_i = 1\})}\right\} = \left\{\frac{\sum_t \hat{w}_t A_{\mathcal{R}_t^{(start)} \cup \mathcal{Y}_t^{(start)}}}{\sum_t \hat{w}_t(\#\{i \in \mathcal{R}_t^{(start)} \cup \mathcal{Y}_t^{(start)} : z_i = 1\})}\right\},$$

against a \$>\$ alternative, here with conditioning on the assignment vector $\mathbf{z}$. This is a rough equivalent to testing the null of

$$\sum_t \bar{w}_t \operatorname{E}(Y_T - Y_C | \mathcal{G}_t^{(start)}, Z = 1) = \sum_t \bar{w}_t \operatorname{E}\{Y_T - Y_C | (\mathcal{R}_t^{(start)} \cup \mathcal{Y}_t^{(start)}), Z = 1\},$$

against the alternative that

$$\sum_t \bar{w}_t \operatorname{E}(Y_T - Y_C | \mathcal{G}_t^{(start)}, Z = 1) > \sum_t \bar{w}_t \operatorname{E}\{Y_T - Y_C | (\mathcal{R}_t^{(start)} \cup \mathcal{Y}_t^{(start)}), Z = 1\}.$$

Also H_3' refers to a test of the same null against the opposite alternative.

Each test H_k serves as a gatekeeper for the test of H_k': H_k' is not tested unless H_k is significant at level .10; if H_k' is tested, it receives a p-value equal to the maximum of its local p-value and the p-value that was attached previously to H_k. In fact, H_k can serve as a gatekeeper not only for the test of whether average effects are equal within the subgroup k and its complement but also for tests of each hypothesis assigning a specific scalar value to that difference, with no further requirement for FWER control, as follos from the argument of Paul R Rosenbaum (2008).

Additional moderation and mediation analyses

Additional moderation and mediation-related hypotheses will be conducted, with p-values adjusted for control of the false discovery rate using the Benjamini-Hochberg method:

1. Effects are greater for students flagged as economically disadvantaged.
2. Effects are lesser for students flagged as economically disadvantaged.
3. Effects are greater for English language learners.
4. Effects are lesser for English language learners.
5. Pupils joining the study as kindergarteners in year 3 experience greater benefit over their 2 years of observation than did pupils joining at grade K in study year 1 over a comparable observation period (i.e. first 2 years of study).
6. Effects of Burst:Reading associate positively with the regression of control-condition early reading achievement on baseline characteristics of students and schools, i.e. with predicted values emanating from the Peters-Belson fit.
7. Effects of Burst:Reading associate negatively with the regression of control-condition early reading achievement on baseline characteristics of students and schools, i.e. with predicted values emanating from the Peters-Belson fit.
8. After fitting to treatment group schools a regression model predicting levels of their implementation from baseline measures, the linear combination of baseline variables used by the regression model to predict implementation associates positively with school-specific Peters-Belson effect parameters, i.e.

   $$\sum_{i} \bar{w}_i \{\#(\mathcal{S} \cap \mathcal{F}_i)\}^{-1} \sum_{i \in \mathcal{S} \cap \mathcal{F}_i} (y_{Ti} - \hat{\mu}_{Ci}),$$

   where $\mathcal{S}$ is the collection of observations on all students joining the study at a given school and y refers to the Star Early Learning outcome.
9. As part of a related but separate study by members of the same study team, a model was fit to data on a separate sample of Burst-participating schools that predicted their average spring DIBELS scores on the basis of fall DIBELS scores and school-level covariates. The fitted values of this model associate positively with school-specific Peters-Belson effect parameters.
10. A model was fit to data on a separate sample of schools, predict fidelity/intensity of program implementation on the basis of baseline school variables. Applying the fitted model to baseline variables from the BURST RCT sample gives rise to in implementation index. This index associates positively with school-specific Peters-Belson effect parameters.
11. A model was fit to the universe of K-3 serving schools listed in the CCD, predicting whether a school subscribes to the Burst service on the basis of school-level characteristics. The participation propensities from this model associate positively with school-specific Peters-Belson effect parameters.
12. The model of implemention that was fit to the treatment group also fitted per-school random effects. These random effects associate positively with school-specific Peters-Belson effect parameters. (A hypothesis about implementation fidelity.)
13. The model of implemention that was fit to the treatment group also fitted per-student random effects, in addition to per-school random effects. Among treatment group students, school-specific Peters-Belson effect parameters associate positively with sums of school and student random effects from the implementation model. (These sums being interpretable as the number of Burst cycles the student received minus the number of Burst cycles that a student with his baseline characteristics would be expected to receive. This is a student-level dosage hypothesis.)

Tests (6) and below seek evidence of association between variables constructed during analysis, for example Peters-Belson effect parameters. These variables involve estimated coefficients. Each of the corresponding statistical hypotheses under test asserts an association between the version of the constructed variable involving (randomization-based) expected values of those coefficients, not the sample-based, random coefficients. To test these hypotheses we'll bootstrap Kendall's τ statistic, with re-fitting of coefficients incorporated into bootstrap iterates.

References

Belson, William A. 1956. "A Technique for Studying the Effects of a Television Broadcast." *Applied Statistics* 5 (3): 195–202.

Benjamini, Yoav, and Yosef Hochberg. 1995. "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing." *Journal of the Royal Statistical Society. Series B (Methodological)*. JSTOR, 289–300.

Benjamini, Yoav, and Daniel Yekutieli. 2001. "The Control of the False Discovery Rate in Multiple Testing Under Dependency." *Annals of Statistics*. JSTOR, 1165–88.

Cochran, W. G. 1969. "The Use of Covariance in Observational Studies." *Applied Statistics* 18: 270–75.

Efron, Bradley. 2012. *Large-Scale Inference: Empirical Bayes Methods for Estimation, Testing, and Prediction*. Cambridge University Press.

Hansen, Ben B., and Jake Bowers. 2009. "Attributing Effects to a Cluster Randomized Get-Out-the-Vote Campaign." *Journal of the American Statistical Association* 104 (487): 873–85.

Holland, P. W. 1986. "Statistics and Causal Inference (with Discussion)." *Journal of the American Statistical Association* 81: 945–70.

Hothorn, Torsten, Frank Bretz, and Peter Westfall. 2008. "Simultaneous Inference in General Parametric Models." *Biometrical Journal* 50 (3). Wiley Online Library: 346–63.

Neyman, J. 1923. "On the Application of Probability Theory to Agricultural Experiments. Essay on Principles. Section 9." *Statistical Science* 5: 463–80.

Pashley, Nicole E, and Luke W Miratrix. 2017. "Insights on Variance Estimation for Blocked and Matched Pairs Designs." *arXiv Preprint arXiv:1710.10342*.

Peters, Charles C. 1941. "A Method of Matching Groups for Experiment with No Loss of Population." *Journal of Educational Research* 34: 606–12.

Pustejovsky, James E, and Elizabeth Tipton. 2017. "Small-Sample Methods for Cluster-Robust Variance Estimation and Hypothesis Testing in Fixed Effects Models." *Journal of Business & Economic Statistics*. doi:10.1080/07350015.2016.1247004.

Romano, Joseph P, and Michael Wolf. 2005. "Exact and Approximate Stepdown Methods for Multiple Hypothesis Testing." *Journal of the American Statistical Association* 100 (469). Taylor & Francis: 94–108.

Rosenbaum, Paul R. 2008. "Testing Hypotheses in Order." *Biometrika*. Biometrika Trust.

Rosenbaum, Paul R. 2001. "Effects Attributable to Treatment: Inference in Experiments and Observational Studies with a Discrete Pivot." *Biometrika* 88 (1): 219–31.

Rubin, D. B. 1974. "Estimating the Causal Effects of Treatments in Randomized and Nonrandomized Studies." *J. Educ. Psych.* 66: 688–701.

Schochet, Peter Z. 2008. "Technical Methods Report: Guidelines for Multiple Testing in Impact Evaluations. Ncee 2008-4018." *National Center for Education Evaluation and Regional Assistance*. ERIC.

Wyss, R., B. B. Hansen, A. R. Ellis, J. J. Gagne, R. J. Desai, R. J. Glynn, and T. Stürmer. 2017. "The 'Dry-Run' Analysis: A Method for Evaluating Risk Scores for Confounding Control." *American Journal of Epidemiology* 185 (9): 842–52. doi:10.1093/aje/kwx032.

Appendix B

Implementation Prognostic Score

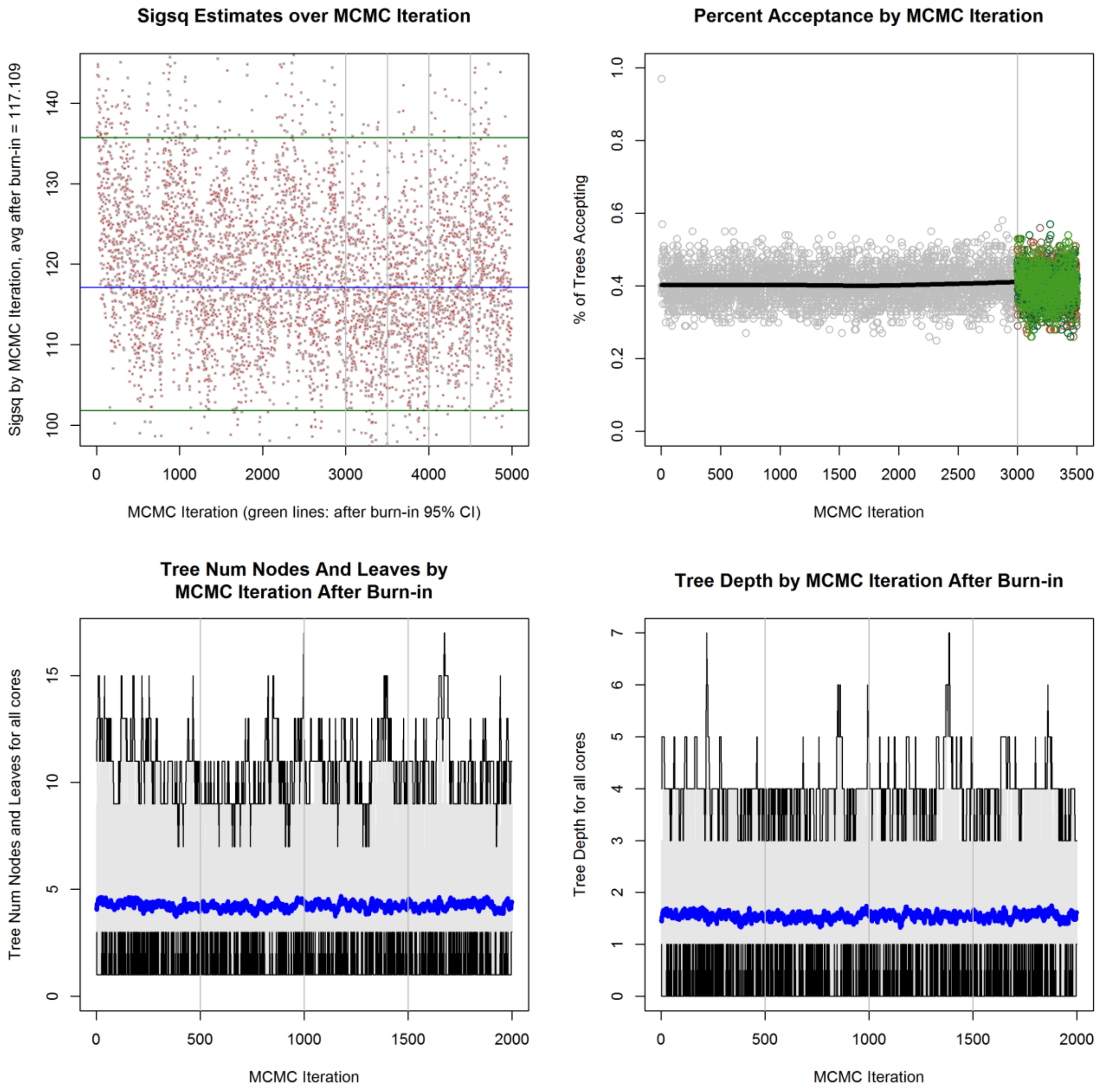
The goal of this appendix is to provide more detailed information on the implementation prognosis score. The implementation prognosis score was intended to predict a school's typical usage of the Burst program, based on the population of all Burst users contained in the M-CLASS data system. After training the model in the population of Burst users, the model was used to predict implementation across all schools. Because the goal was prediction accuracy, we chose the machine-learning algorithm Bayesian Additive Regression Trees (BART; Kapelner & Bleich, 2013), which shows good prediction quality across a range of different problems. BART models are a type of sum-of-trees ensemble method using fully Bayesian probability models. We used 100 trees and other default function parameters (after checking with cross-validation that these were optimal). BART models were run in R using the bartMachine package. More information about these models can be found in Kapelner & Bleich (2013).[12] The following variables were used as predictors: SEDA: Community Socio-Economic Status, SEDA: Segregation Index, SEDA: Cohort Adjusted Growth in ELA, Number of Students in District, Indicator for School Title I Status, Magnet School Indicator, Charter School Indicator, indicators for whether the school is Rural, in a Town, in a City, or in a Suburb, Number of Students, Size of Pre-K, Pct Students on FRL, Pct Hispanic Students, Pct Black Students, Total Per-Pupil District Expenditure, Standardized Pct 3rd Graders Proficient Across Reading and Math, 1 year Lagged Standardized Pct 3rd Graders Proficient Across Reading and Math. SEDA indicates variables are community/district level variables from the Stanford Education Data Archives (Reardon, Ho, Shear, Fahle, Kalogrides, & DiSalvo, 2017).

Description of the model fit is below:

```
## bartMachine v1.2.3 for regression
##
## Missing data feature ON
## training data n = 620 and p = 23
## built in 13.3 secs on 4 cores, 100 trees, 3000 burn-in and 2000 post. samp
les
##
## sigsq est for y beforehand: 171.855
## avg sigsq estimate after burn-in: 117.10898
##
## in-sample statistics:
##  L1 = 4140.02
##  L2 = 58950.27
##  rmse = 9.75
##  Pseudo-Rsq = 0.5851
## p-val for shapiro-wilk test of normality of residuals: 0
## p-val for zero-mean noise: 0.726
```

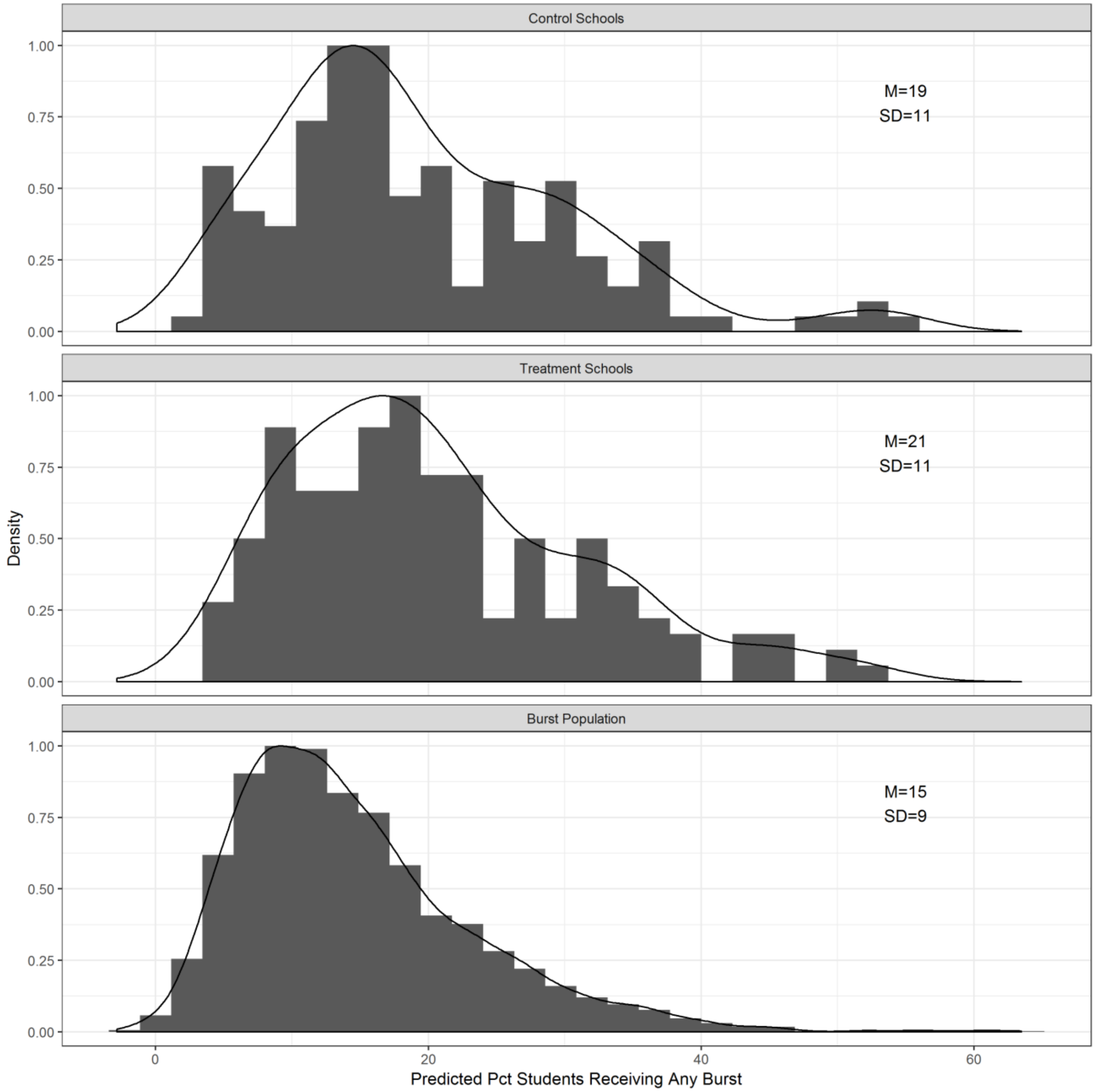
The model is fit using Markov-Chain Monte-Carlo (MCMC) with 3,000 burn-in samples and 2,000 post-burn-in samples. The variance in the percentage of students receiving any Burst cycles across schools is 117 after model convergence, giving a pseudo-R^2 of 0.59. Test for residual normality show that residuals are not normally distributed, which would call into question confident intervals (we, however, do not use confidence intervals). Convergence diagnostics are shown in the graph below. The top-left graph shows variance estimates across MCMC iterations. To the left of the first grey vertical line is the burn-in sample (for chain 1), and each

[12] Kapelner, A., & Bleich, J. (2013). bartMachine: Machine Learning with Bayesian Additive Regression Trees. *ArXiv:1312.2171 [Cs, Stat]*. Retrieved from http://arxiv.org/abs/1312.2171



of the four areas after the first grey line are the post-burn-in samples across the four chains. While there is little initial improvement in residual variance estimates, suggesting relatively little learning after setting initial model parameters, the chains all converge to the same residual variance estimate. The top-right graph shows the percentage of generated trees accepted by the algorithm, with different colors after the grey line indicating different chains. Again, we see convergence at a relatively high acceptance rate of 40%. The bottom two graphs show the number of leaves on each tree and tree depth across MCMC iterations, with only post-burn-in samples displayed and with each box (as indicated by grey vertical line) showing a different chain. The blue line shows averages across trees at each step and the black lines shown min/max values. These also show convergence across chains.

The graphs on the next page show the distribution of the implementation prognosis score (i.e. predicted percentage of students with receiving any Burst). In the Burst population, there is positive skew, but the distribution is roughly normal, centered on a mean of 15% (SD=9). The distributions for the control and treatment



schools are roughly the same, but the means are about 5 percentage points higher and the standard deviations are higher. The means and variances of the distributions are not significantly different between the treatment and control schools, but both RCT groups have significantly higher means ($t=4.8$, $p<0.01$ and $t=6.0$, $p<0.01$ for the control and treatment schools respectively) and standard deviations ($F=0.64$, $p<0.01$ and $F=0.66$, $p<0.01$ for the control and treatment schools respectively) than the Burst population. As we show later, this is due to the sample of schools in the RCT being the type of school that happens to implement the RCT at higher rates, rather than different patterns of implementation in the experiment versus the Burst population schools.

Table B1: Comparison of All Schools, Burst Users, and RCT schools in 2013

Variable	Population Frame	Burst Users	Ever Recruited	Breakdown of RCT Schools Attriters	Retained
School Characteristics					
Title I School	0.80	0.78	0.97	0.95	0.98
School in City	0.28	0.36	0.37	0.53	0.25
School in Town/Rural	0.38	0.34	0.49	0.45	0.52
School in Suburb	0.33	0.31	0.14	0.03	0.23
Teacher -Child Ratio	0.07	0.06	0.06	0.07	0.06
Number of Students	468	466	401	419	387
Pct White Students	0.54	0.48	0.50	0.38	0.59
Pct Students on FRL	0.55	0.58	0.73	0.70	0.75
Student Achievement					
Std Pct Proficient in 3rd Reading	0.03	-0.02	-0.21	-0.18	-0.23
Std Pct Proficient in 3rd Math	0.03	-0.01	-0.31	-0.15	-0.44
Missing Values					
Teacher -Child Ratio	0.03	0.00	0.00	0.00	0.00
Std Pct Proficient in 3rd Reading	0.16	0.11	0.18	0.25	0.13
Std Pct Proficient in 3rd Math	0.16	0.11	0.16	0.25	0.10

M-CLASS Data

Amplify provided us a data set from their M-Class computer system that tracks Burst usage for all schools using Burst. This data was initially provided in December of 2017 and provided again with NCES IDs in June of 2018. The data was for 2016-2017 and included all 651 schools that were using Burst during this school year. The data included semester level information, including the number of students at each Tier who received any Burst cycles and the number of students who received 6 or more Burst cycles each semester. Data on the number of students who took DIBELS as well as DIBELS school averages for the beginning, middle, and end of year time points were also provided. This data was linked to the CCD database to get school level characteristics for the schools.

Create Comparison of Populations

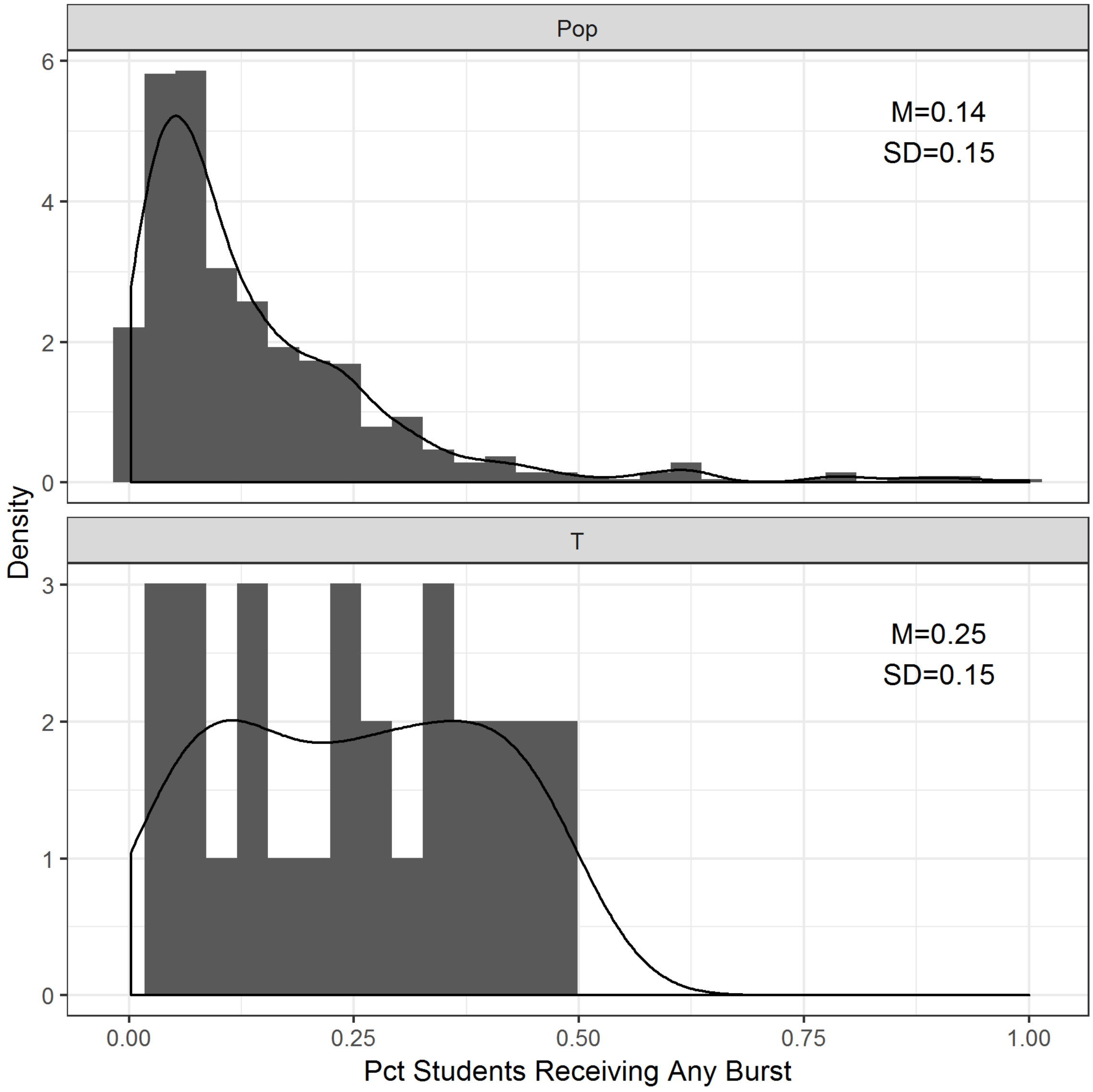
The population frame is created from all schools in the Common Core of Data that offer any grades K-3, in one of the 50 states or DC, labeled as a "Regular School" and not closed or inactive, and listed as having at least one student. Table B1 (above) shows differences between the Population frame, the Burst Users, and the RCT schools. While Burst Users were in poorer communities than the frame as a whole, the RCT schools were in much poorer communities than the Burst Users. It is then, no surprise that RCT schools are more likely to be Title I and less likely to be in the suburbs than either the frame or Burst Users. Burst Users seem to be in much larger districts than average while the RCT recruited from smaller districts than average. As was an intentional aspect of RCT sample recruitment, the RCT schools have higher percentages of students on Free-Reduced Price Lunch (FRL) and more Hispanic (but not more Black) students than the frame. Last, RCT schools are much lower achieving than either the frame or the Burst users.

Table 2: Comparison of All Schools, Burst Users, and RCT schools in 2013 on Variables Not Used in Report

				Breakdown of RCT Schools	
Variable	Population Frame	Burst Users	Ever Recruited	Attriters	Retained
SEDA: Community Socio-Economic Status	-0.03	-0.15	-0.52	-0.30	-0.68
Number of Students in District	16,366	45,202	8,315	8,191	8,410
Schoolwide Title I	0.54	0.61	0.88	0.77	0.96
Targeted Title I	0.14	0.06	0.08	0.18	0.00
Pct Hispanic Students	0.22	0.27	0.29	0.41	0.19
Pct Black Students	0.15	0.18	0.15	0.14	0.17

SEDA indicates variables from the Stanford Education Data Archives. SEDA Segregation Index averages the SEDA segregation indexes across FRL and Race.

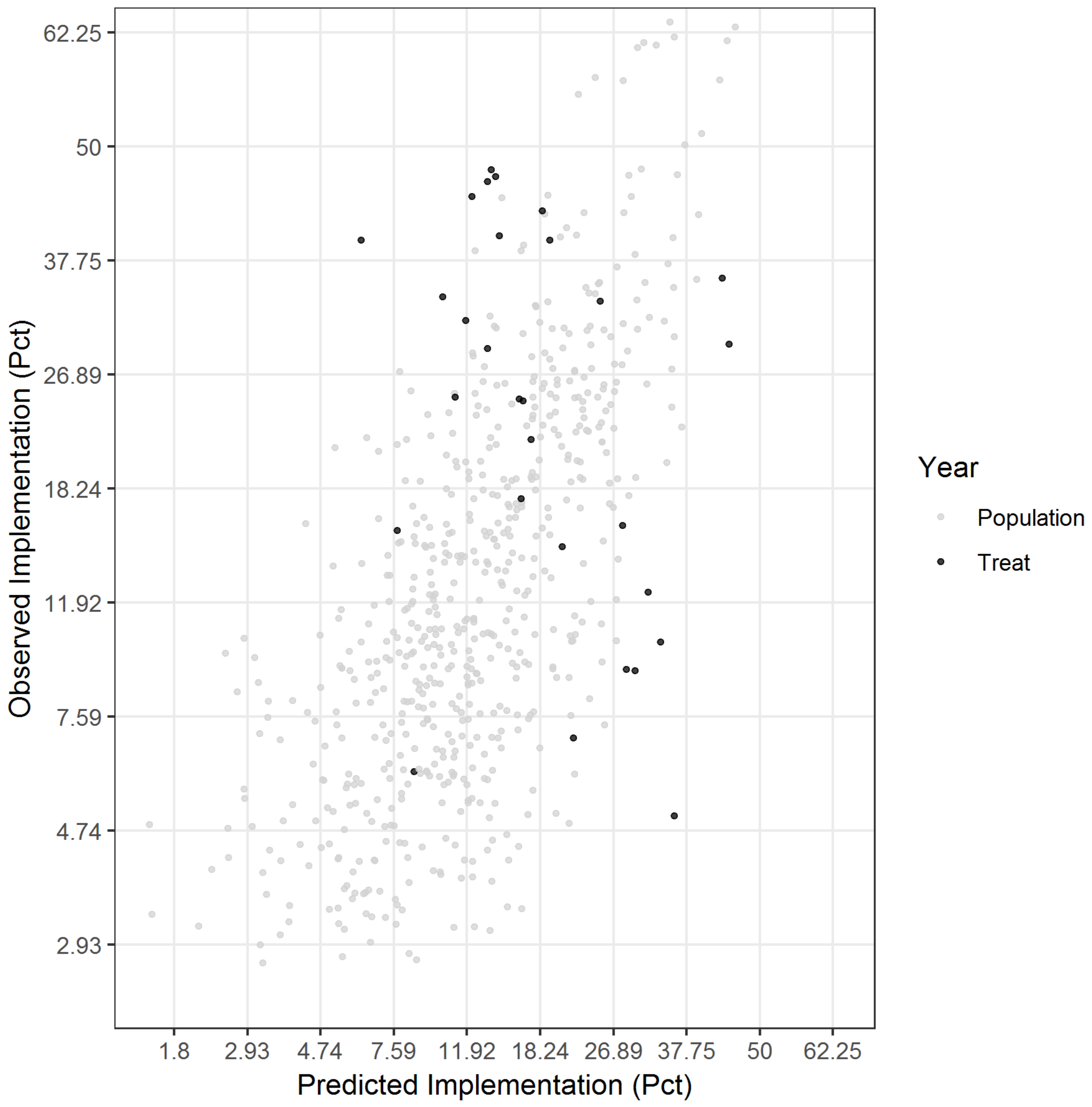
Predicted Versus Observed Implementation



```
##
##  Welch Two Sample t-test
##
## data:  Pct_Any by TreatITT
## t = -3.8, df = 31, p-value = 0.0006
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.1630 -0.0492
## sample estimates:
## mean in group Pop   mean in group T
##            0.1442            0.2503
```

```
##
##  F test to compare two variances
##
## data:  Pct_Any by TreatITT
## F = 1.1, num df = 620, denom df = 28, p-value = 0.9
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.5786 1.7201
## sample estimates:
## ratio of variances
##              1.067
```

We also tested for differences in mean and variance across the two samples. The RCT schools have a significantly larger percentage of students receiving any Burst cycles (t =-3.8, df = 31, p-value = 0.0006). The differences in variance are not significant (F = 1.1, num df = 620, denom df = 28, p-value = 0.9).



The graph above shows the observed percentage of students receiving any Burst cycles in a semester (aggregated across semesters and years) versus the implementation prognosis score (a predicted value for this variable). This prediction comes from a machine-learning Bayesian Additive Regression Tree model trained on the M-CLASS data and used to generate predicted values for M-CLASS and RCT schools. While the axes labels portray the percentage of students, the axes are on a logit scale, which spreads out values at the lower end of the scale, allowing for a better visual presentation. Note that some schools in the population have 100% of students receiving some Burst and these schools are not shown on the graph, which was scaled to show the Treatment Schools. This graph highlights that the implementation rates, which are higher in the RCT, are driven by the fact that RCT schools are in a subset of schools with higher expected rates of implementation. That is, they are further to the right on the graph that the average population school. Implementation rates in treatment schools actually look quite similar to population schools when accounting implementation prognosis. That is, the black dots fall roughly within the area where the grey dots exist.

Appendix C

Report on Professional Development Survey for Teachers

Methods

Sample. The results presented in this report were obtained from the 2016-2017 Professional Development Survey (PDS) administered to 666 teachers in both treatment and control schools for the BURST study. Amplify produced a roster of 580 teachers for the study. The roster constituted the sampling frame for the study. 11 of the teachers from this roster were no longer active members of the treatment or control schools and were dropped from the roster, leaving a roster of 569 teachers. This roster was then merged with the PDS data, yielding a total of 543 teachers from the Professional Development Survey data who were also on the roster.

The PDS consisted of 56 questions concerning each teacher's professional development experiences concerning reading. 134 teachers failed to answer any questions of the survey, yielding a group of 409 teachers who answered the PDS. Selected questions from the PDS were combined to yield 6 scales. The compositions of those scales is given in Table 1.

Data Analysis

Our analyses sought to answer 3 questions.

1. Did treatment impact whether teachers completed the PDS?
2. Did treatment impact teachers' self-reported professional development?
3. How much of the variability in teachers self-reported professional development was due to school effects
4.

The first question was addressed by a cross-tabulation table of PDS completion by treatment group with a Pearson Chi-square test of independence for the framed sample of 543 teachers.
The second question and third questions were addressed by a hierarchical linear model/mixed model analysis of the 6 constructed scales for the PDS. Question 2 was addressed by the fixed effect of treatment in the mixed model analysis while question 3 was addressed by the estimated random effects of school in the mixed model analysis.

Results

A cross-tabulation table for the teachers who completed the PDS by treatment group is given in Table 2. One out of the 409 teachers who completed the PDS was not in a treatment or control school and is omitted from this table[13]. The Pearson Chi-Square test of independence yielded a value of 0.57 and was not statistically significant (p-value = 0.45)

The results for the mixed model analysis of the 6 PDS scales is given in Table 3. Only 1 of the 6 estimated treatment effects was greater in absolute value than its associated standard error. None of the 6 treatment effects were statistically significant; p-values that ranged from 0.17 to 0.97, with only 1 p-value less than 0.4. Generally school effects explained only a small proportion of the variance for each of the 6 PSD scales. Proportion of variance explained by school effects ranged from 0% to 15.5%. Only 1 of the variance proportions was greater than 10%, with 3 out of the 6 proportions less than 5%.

[13] That teacher from Holly District had school listed as Other

Table 1: Composition of Professional Development Scales

1. Professional Development in Reading
Attended short, stand-alone training or workshop in reading (half-day or less)
Attended longer institute or workshop in reading (more than half day)
Attended a college course about reading
2. Professional Development in Reading Assessment
Administering skills-based formative assessments
Administering another type of formatives assessment in reading
Interpretation of results of reading assessments
3. Professional Development in Teaching Practices of Reading
Delivering reading intervention to struggling readers
Grouping students for reading instruction
Learning the pedagogy of different instructional approaches
4. Professional Development on Phonics
Teaching the alphabetic principle
Teaching fluency
Teaching phonemic awareness
Teaching phonics
5. Professional Development for On-Site Practices
I observed demonstrations of teaching techniques
I practiced what I learned and received feedback
I participated in and/or led group discussions
I observed or conducted a demonstration of a lesson, unit or skill
I developed and practiced using student materials
6. Professional Development Improved My Knowledge and Skills
Improved my knowledge of how children learn to read
Improved my knowledge of the early grades reading and language arts curriculum
Improved my classroom management skills
Improved my understanding of effective instructional strategies for teaching reading
Improved my ability to teach reading to diverse student populations
Improved how I assess student learning
Improved how I group students for instruction

Table 2: Cross-Tabulation of PD Survey by Group

	Completed PD Survey		
Group	Yes	No	Total
Control Count	195	59	254
% within Control	76.77%	23.23%	100.00%
Treatment Count	213	75	288
% within Treatment	73.96%	26.04%	100.00%
Total Count	408	134	542
% within Total	75.28%	24.72%	100.00%

Pearson Chi-Square = 0.57 p = 0.45

Table 3: Mixed Model Results for Professional Development Survey Scales

	PD Reading		PD Assessment		PD Teaching Practices	
Response	Estimate	S.E.	Estimate	S.E.	Estimate	S.E.
Fixed Part						
constant	12.24	1.05	4.61	0.27	4.50	0.25
Treatment	2.00	1.45	-0.18	0.37	0.13	0.34
Random Part						
School Variance	0.00	0.00	0.61	0.34	0.28	0.28
Teacher Variance	215.14	15.06	8.12	0.61	8.81	0.66
School Variance %	0.00%		7.01%		3.04%	
	PD Phonics		PD Onsite Practices		PD Improve Skills	
Response	Estimate	S.E.	Estimate	S.E.	Estimate	S.E.
Fixed Part						
Constant	5.10	0.34	6.30	0.33	21.85	0.76
Treatment	0.13	0.47	0.35	0.46	0.04	1.04
Random Part						
School Variance	0.43	0.53	1.10	0.52	7.54	2.69
Teacher Variance	17.83	1.34	10.93	0.83	41.11	3.12
School Variance %	2.37%		9.16%		15.49%	