Ready for Fall?

Near-Term Effects of Voluntary Summer Learning Programs on Low-Income Students' Learning Opportunities and Outcomes

TECHNICAL APPENDIXES

Jennifer Sloan McCombs, John F. Pane, Catherine H. Augustine, Heather L. Schwartz, Paco Martorell, and Laura Zakaras

Commissioned by

The Wallace Foundation

Supporting ideas.
Sharing solutions.
Expanding opportunities.

RAND CORPORATION

For more information on this publication, visit www.rand.org/t/rr815

Published by the RAND Corporation, Santa Monica, Calif.

Table of Contents

Figures and Tables

Figures

Tables

Abbreviations

DESSA-RRE	Devereux Student Strengths Assessment—RAND Research Edition
ELL	English language learner
FRPL	free and reduced-price lunch
GAMM	Generalized Additive Mixed Models
ITT	intent-to-treat
MDES	minimum effect size the study would be able to detect
OLS	ordinary least squares
TOT	treatment effect on the treated
WWC	What Works Clearinghouse

Appendix A

Randomization Design and Implementation

In this appendix, we discuss details of how randomization was conducted, power for detecting treatment effects, attrition, and the comparability of the treatment and control groups after attrition.

Randomization of Students to Treatment and Control Groups

In each district, students applied to participate in the summer program, but the specific site to which they would be assigned (if admitted) was determined by the district, typically through geographic feeder patterns.

Stratification Plan

Our thinking about how to design the experiment was strongly influenced by a recent paper by Guido Imbens (2011) that discusses the methodological considerations when designing a randomized controlled trial experiment. He shows that partitioning the sample into strata (groups of individuals with similar characteristics) and then randomizing within strata is preferable to randomizing without first creating strata, from the standpoint of maximizing statistical power, and that the benefits of randomization are strongest when (1) there are "relatively many and small strata" and (2) the strata are based on covariates that are strongly related to the outcome of interest.

With these considerations in mind, we constructed strata for the experiment based on the following variables:

- district
- third-grade school
- English language learner (ELL) status in third grade
- race: Hispanic, black, other[1]
- eligibility for either free or reduced-price lunch (FRPL) in third grade
- prior achievement.

We chose these variables for several reasons. First, these are variables that each district in the study maintains. Second, these are all factors that are well known to be strong predictors of student achievement. Finally, variables like ELL are related to the type of test a student takes or whether a student is tested at all. As we will discuss, we conducted robustness checks where we limited the sample to particular student subgroups, including students tested only in English,

[1] For the purposes of the stratification, "Hispanic" will refer to students of Hispanic origin of any race and "black" will refer to non-Hispanic black students.

since the testing conditions may not be comparable for ELL students. Since some researchers advocate for stratifying on variables that will be used in subgroup analyses (Duflo, Glennerster, and Kremer, 2007), we thought it important to include these factors in the definition of the randomization strata.

An important issue for the stratification was what achievement test to use. In some districts, systematic testing begins in spring of the second grade, whereas other districts begin with benchmark assessments in the fall of third grade. Further, students are tested in multiple subjects. We stratified on the most-recent test scores that were available for all students. In districts where all students were tested in reading and mathematics, we used the average of these two subjects as the stratifying variable. A different approach was used in sites that administered some of their tests in Spanish to Spanish-speaking ELL students. For example, in Dallas, the stratification was based on mathematics tests, which are administered in English to all students.

While Imbens (2011) argues in favor of using strata with relatively few students, he also points out that there are analytic challenges associated with having very few (e.g., two) students per strata.[2] A practical limitation of having strata with too few observations is that if students drop out of the sample due to attrition and only control or treatment students are left, the entire strata must be dropped from the analysis. Relatedly, when conducting "treatment effect on the treated" (TOT) analysis, if there is no effect of the randomization on the probability of attending summer school for a particular strata (e.g., only no-shows among the students assigned to treatment), the entire strata will not contribute to the estimate of the treatment effect. Thus, we defined the strata to have about 15–30 students each. For example, with ten students assigned to treatment in a stratum, it was very unlikely that all of the students would attrit from the study or be treatment no-shows.[3]

The basic algorithm used to generate the strata was as follows:

1. We fully stratified the summer program enrollees into cells defined by district-school-race-ELL-FRPL-achievement. For this first step, achievement is a binary variable with half the enrollees in a district in a high-achievement group and the other half in a low-achievement group.[4] Since a goal of the study was to examine effects by higher- and lower-achieving students, it was important to stratify the sample in this way so that the high/low achievement subgroups coincided with the strata.
2. For cells defined in Step 1 with at least one but fewer than 15 students, we aggregated cells until each had at least 15 students. Cells were aggregated in the following order: (a)

[2] Specifically, it was impossible to compute the variance of the within-strata treatment effect estimate. The estimated variance of the treatment effect in the case with just two observations per strata (i.e., a paired randomization design) would be biased upward for the true variance of the estimated treatment effect.

[3] Suppose that five students dropped out of the study (which would be higher than the long-run attrition rate of 30 percent that we assumed). If the no-show rate were 30 percent, then the probability that all five of the non-dropouts were no-shows would be only about 0.002.

[4] Students with missing baseline achievement data were assigned to the low-achievement strata. See our discussion for additional sensitivity checks that were performed to handle missing baseline data.

FRPL, (b) race, (c) ELL, and (d) binary achievement. We did not collapse schools, to ensure that within each school the proportion assigned to treatment was near the intended proportion (as close as possible considering rounding). To see how this worked, suppose that the cell for black students who are FRPL, ELL, low-achievement, and in "School A" contained only five students. The first step in aggregating cells would be to form a cell for students who were in the black or other race categories, FRPL, ELL, low-achievement, and "School A" (i.e., pool across the two smallest race categories). If this new cell still had fewer than 15 students, then the next step would be to pool again by race, and if further pooling were necessary, to then pool by FRPL to form a cell for students who were ELL, low-achievement, and in "School A." This process was repeated until students were all assigned to cells with at least 15 students or to cells at the school level that could not be further aggregated.

The ordering of covariates for the aggregation reflected how important we thought each variable was in the stratification; we aggregated on less-important variables first. We placed the greatest importance on the sending school for programmatic reasons: stratifying on the school ensured that no sending school would have a disproportionately large or small proportion of students assigned to treatment, helping to make clear that all stakeholders were being treated fairly by the randomization process. The second tier of importance was given to the achievement and ELL variables because we examined treatment effects by subgroups according to these variables. Finally, FRPL and race were included in the stratification plan because they were strongly associated with student outcomes; however, they received the lowest priority.

3. For cells that, at the end of Step 2, contained more than 30 students, we stratified them further by the achievement variable to form as many cells as possible that contained at least 15 students. For instance, if a cell had 65 students at the end of Step 2, we formed three cells with 16 students each and one with 17 students.[5] We used achievement to further stratify the larger cells because prior achievement was the strongest predictor of future achievement, so forming strata in this way maximized statistical power.

Writing the Computer Code for the Randomization

We used STATA.do files to carry out the stratified random assignment. We assigned percentage P of student applicants to the treatment group. P varied across districts and summer sites within a district, ranging between 50 and 60 percent. We capped P within any summer site at 70 percent.

Within each strata, P percent of students was assigned to the treatment group. For strata where the number of observations did not enable exactly P percent to be assigned to treatment,

[5] More formally, for a cell, c, with N_c students, we formed $k=int(N_c/15)$ subcells, with $int(N_c/k)$ students in k–mod(N_c,k) cells and $int(N_c/k)$+1 students in mod(N_c , k) cells.

the number of students assigned to treatment was equal to *round(P*N_c)*, where N_c was the number of students in strata *c*. In this way, whether there was slightly more or less than *P* percent of students assigned to treatment varied across strata but this variation was random.

To do the actual random assignment, we used the STATA pseudo-random number generator. Specifically, we assigned each student a random number drawn from a uniform distribution on the interval (0,1). Students were sorted according to this number so that the sort order of the students was random. After sorting the data in this way, within each strata, students whose sort order was less than or equal to N_{tc} were assigned to treatment, where N_{tc} is the number of students in strata *c* assigned to treatment. The strata identifier, *c*, was stored with each record for use in future analyses.

Siblings

It could be disruptive to families if one or more of their children were admitted to the program and one or more were not admitted. In all districts that requested we account for siblings, we adopted procedures to keep together all siblings that made valid applications to the program.[6] Where one of the siblings was in third grade and the remaining siblings were in other grades, the admission decision for all of the children was based on the third-grader's randomly assigned admission status. Where there were multiple siblings in the third-grade sample, the siblings were randomly assigned as a group so that all received admission or all were denied admission.

Program Uptake

Here we discuss estimates of the impact of the randomized treatment assignment on participation in the summer program, which we define as attending at least one day of the summer program. With perfect compliance to the experimental protocol, treatment assignment and program uptake would be the same. However, as Table A.1 indicates, not all students admitted to the summer programs actually attended, and similarly, some students assigned to the control group attended the program.[7]

Table A.1: Impact of Treatment Assignment on the Likelihood of Program Participation

Program Uptake Among Students Assigned to Treatment Group	Program Uptake Among Students Assigned to Control Group	Difference in Uptake
0.799	0.047	0.752

[6] Siblings were not considered when randomizing for Boston, by the decision of the district.

[7] The uptake rates by treatment assignment status were virtually identical for the sample of mathematics and language arts non-attriters. In Dallas, some students assigned to the control group were nonetheless admitted to the program. There were a handful of other such "crossover" control group students in other districts as well.

Table A.2 shows linear probability model estimates of the impact of treatment assignment on program uptake. Standard errors were calculated using the Eicker-Huber-White sandwich estimator that is robust to heteroskedasticity (Eicker, 1967; Huber, 1967; White, 1980). The results indicate that assignment to be eligible for the program increased the likelihood of attending the program for at least one day by 75 percentage points. Where we present TOT estimates, we report instrumental variable estimates of the impact of summer program attendance for the set of students whose summer program attendance was affected by the experimental assignment, which are equal to the intent-to-treat (ITT) estimates scaled up by the inverse of the estimates in Table A.2.

Table A.2: Impact of Treatment Assignment on the Likelihood of Program Participation, by Subject

	Estimate	Standard Error	p-value
Mathematics non-attriters, only strata fixed effects	0.751	0.009	0.000
Mathematics non-attriters, all student-level covariates	0.751	0.009	0.000
Reading non-attriters, only strata fixed effects	0.751	0.009	0.000
Reading non-attriters, all student-level covariates	0.751	0.009	0.000

NOTE: Student-level covariates are standardized mathematics and reading scores from the state's spring third-grade assessments and fall third-grade diagnostic tests, classroom average of these pretests, dummy variables for FRPL, black, Hispanic, ELL, special education, male, and missing pretest score dummy variables.

Minimum Detectable Effect Sizes

To estimate the statistical power of the study during the design phase, we applied formulas for research experiments in which treatment students were clustered in classes, and calculated the minimum effect size the study would be able to detect (MDES) with 80 percent probability using a two-tailed test and a 0.05 level of significance. To perform these calculations, we estimated several parameters using existing empirical data from pilot work in summers 2011 and 2012 as well as the published literature. These estimates were uncertain, and as a general rule we chose conservative values that would produce higher, rather than lower, MDES estimates.

Once the student sample and the proportion assigned to treatment were final, we used this information, along with the remaining assumptions from the earlier power calculations, to compute the MDES for near-term intent-to-treat outcomes. These MDES values are shown in Table A.3 for the overall study as well as district-specific descriptives.

Table A.3: Minimum Detectable Effect Sizes for Intent-to-Treat Analyses of Near-Term Student Outcomes

	MDES
Overall	0.08
Boston	0.19
Dallas	0.13
Duval	0.20
Pittsburgh	0.23
Rochester	0.18

Attrition

There were 5,639 students in the initial sample for the experiment. However, outcomes were not available for all students because of withdrawal from the study, refusal to take the fall 2013 tests, prolonged school absence during the fall 2013 test administration window, student mobility during the test administration, or other reasons a student might have had "unknown" status in the school district during the fall 2013 test administration window. As a result, mathematics outcomes were available for 5,127 students and reading outcomes were available for 5,101 students. These represent attrition rates of 9.1 percent and 9.5 percent, respectively.

To test whether there was differential attrition in the treatment and control groups, we ran linear probability models that predicted attrition based on the treatment indicator. The first model included fixed effects for random assignment strata but no other covariates; the second also included the student-level covariates already discussed. Standard errors were calculated using the Huber-Eicker-White sandwich estimator that is robust to heteroskedasticity. The results show that treatment group students had a slightly lower tendency to attrit, but the results were not statistically significant (see Table A.4).

Table A.4: Assessment of Differential Attrition

	Estimate	Standard Error	p-value
Mathematics, only strata fixed effects	–0.012	0.008	0.132
Mathematics, all student-level covariates	–0.012	0.008	0.139
Reading, only strata fixed effects	–0.009	0.008	0.241
Reading, all student-level covariates	–0.006	0.008	0.422

NOTE: Student-level covariates are standardized mathematics and reading scores from the state's spring third-grade assessments and fall third-grade diagnostic tests, classroom average of these pretests, dummy variables for FRPL, black, Hispanic, ELL, special education, male, and missing pretest score dummy variables.

Balance of the Treatment and Control Groups After Attrition

Next we assessed balance in the observable characteristics of the treatment and control groups that were retained in the analytic sample after attrition. Table A.5 shows the results for mathematics from statistical models that predicted assignment to the treatment group based on each student-level achievement or demographic variable, fit one at a time, controlling for the strata used in random assignment. Table A.6 shows the corresponding results for reading.

Table A.5: Assessment of Treatment-Control Group Balance After Attrition (Mathematics)

	Estimate	Standard Error	p-value
2012 benchmark mathematics assessment	0.009	0.009	0.353
2013 state mathematics assessment	0.015	0.008	0.061
2012 benchmark reading assessment	0.000	0.009	0.992
2013 state reading state assessment	0.007	0.008	0.341
Eligible for free or reduced-price lunch	0.002	0.026	0.938
Black	0.012	0.019	0.509
Hispanic	0.004	0.022	0.845
English-language learner	-0.007	0.020	0.721
Special education student (gifted excluded)	0.043	0.025	0.088
Student is male	-0.003	0.015	0.821

NOTE: Table shows results of univariate models (with strata fixed effects) using each covariate to predict treatment in the sample remaining after attrition. A likelihood ratio test of these variables' joint ability to predict treatment assignment in a multivariate model yielded a p-value of 0.540.

Table A.6: Assessment of Treatment-Control Group Balance After Attrition (Reading)

	Estimate	Standard Error	p-value
2012 benchmark mathematics assessment	0.011	0.009	0.228
2013 state mathematics assessment	0.018	0.008	0.025
2012 benchmark reading assessment	0.003	0.009	0.760
2013 state reading state assessment	0.010	0.008	0.200
Eligible for free or reduced-price lunch	-0.001	0.026	0.980
Black	0.017	0.019	0.372
Hispanic	-0.001	0.022	0.972
English-language learner	-0.013	0.020	0.505
Special education student (gifted excluded)	0.044	0.025	0.078
Student is male	-0.006	0.015	0.704

NOTE: Table shows results of univariate models (with strata fixed effects) using each covariate to predict treatment in the sample remaining after attrition. A likelihood ratio test of these variables' joint ability to predict treatment assignment in a multivariate model yielded a p-value of 0.322.

In both cases, the differences between the retained treatment and control groups were generally very small and not significant. An exception was that in both mathematics and reading, the treatment group had slightly higher scores on the 2013 state assessment, by 0.015 and 0.018,

respectively. The difference was marginally significant in mathematics and significant at the $p < 0.05$ level in reading. These differences were very small compared to what is considered acceptable in a valid experiment.[8] Moreover, when conducting the 20 statistical tests in these tables with a threshold of significance of 0.05, a statistically significant result could have easily arisen by chance. Finally, when all of these covariates were used together to predict treatment assignment, the result was not significant for either mathematics or reading. As a result of these analyses, we concluded that attrition was not a major concern for the validity of our analyses.

[8] For example, the What Works Clearinghouse (WWC) (U.S. Department of Education, 2014) sets a limit of 0.25 for pretreatment group differences when the variable will be used as a covariate in outcomes models, as we did here.

Appendix B

Statistical Analysis

Analysis Plan

Preferred Random-Effects Model

We estimated two types of regression models for ITT analysis of the impact of the summer learning programs. The first, and preferred model, is a random-effects model:

$$Y_{qispc} = \alpha T_{ispc} + \beta X_{ispc} + \gamma_s + PreTestMean_c * T_{ispc} + \pi_p * T_{ispc} + \mu_c * T_{ispc} + \varepsilon_{ispc}$$

where:

- Y_{qispc} is the standardized post-test score in subject q for student i in strata s in summer site p in summer classroom c, where p and c are defined to be zero for control group students.
- T_{ispc} is a indicator of assignment to the treatment group.
- X_{ispc} is a vector of baseline covariates (see below).
- γ_s are strata fixed-effects (dummy variables).
- $PreTestMean_{qc} * T_{ispc}$ is a vector of mean pretest values of all students who were assigned to the same summer 2013 classroom in subject q regardless of the students' sending school(s) or of the school(s) at which the students enrolled as of fall 2013. This is zero for all control students. There are four classroom means, one for each of the four pretests (spring 2013 mathematics and language arts, and the earlier assessments in mathematics and language arts that were used for stratification).
- $\pi_p * T_{ispc}$ is a random-effect common to all students in summer site p. Note that this is zero for students in the control group.
- $\mu_c * T_{ispc}$ is a random-effect common to all students in summer classroom c. Note that this is zero for students in the control group.
- ε_{ispc} is a residual, the variance of which is allowed to vary by pattern of available pretests, as we will discuss.

The baseline variables in the model are:

- spring 2013 assessment scores (third grade) in mathematics and language arts, standardized within district; interacted with district dummies; missing scores equal to zero
- spring 2012 or fall 2012 benchmark assessment scores in mathematics and language arts, standardized within district; interacted with district dummies; missing scores equal to zero
- dummy variables for patterns of missing values of the pretests by district
- dummy for FRPL
- dummy for black
- dummy for Hispanic

- dummy for ELL
- dummy for special education (exclusive of gifted and talented designation)
- dummy for male.

We ran two versions of the model: one that did not include baseline covariates (i.e., X_{ispc} , $PreTestMean_c$) and one that did. We expected the model with baseline covariates would have a similar point estimate but a much smaller standard error due to the variance reduction from controlling for the pretests.

We estimated this model using STATA's "xtmixed" command and R's "lme" command. We made the assumption that the variance of the random effects was constant. We allowed the variance of the residual error to vary across the patterns of missingness in the pretests by district, because the prior scores were likely strong predictors of outcomes and so the residual variance in outcomes should depend on which and how many prior scores were available.

The Stata command we used was:

> xtmixed [*outcome var*] [*treatment var*] [*baseline variables in X*] [*strata dummies*] [*pretest classroom means*] || [*summer site and classroom ID variables*]: [*summer site and classroom dummies (always zero for the control group)*] , nocons covariance(identity) , reml

The corresponding R command was:

> lme([*outcome var*] ~ [*treatment var*] + [*baseline variables in X*] + [*missing data pattern indicators*] + [*strata dummies*] + [*pretest classroom means*], random = list(dumid = [*summer site and classroom dummies (always zero for the control group)*]),[9] weights=varIdent(form = ~1 | [*missing data pattern indicators*]))

Linear Regression with Cluster Adjustment

A concern with this model is that it requires parametric assumptions about the form of the clustering induced by the program being delivered in summer classrooms (and assumes there is no clustering between any students in the control group). Therefore, we also specified an ordinary least squares (OLS) version of the random-effects model (i.e., the covariates in the model were the same as those in the random-effects model), but we conducted statistical inferences making fewer parametric assumptions and also allowed a more general form of clustering.

To do this, we defined clusters corresponding to each summer site. We then considered each sending school to be a member of exactly one of these summer site clusters, defined as the summer site to which the majority of the school's treatment group students were sent for summer

[9] The random effect specification in the R command is adapted from Lockwood, Doran, and McCaffrey (2003), which describes a flexible approach to specifying groupings, including (in the present situation) the partial nesting of only treatment students in summer sites and classes, and (in future years) the additional complexity of cross-classification that will result from students having different site and class assignments in the second summer of this experiment.

2013. Once the clusters were determined in this manner, all treatment students were assigned to the cluster where they actually attended during summer 2013. All control group students (as well as nonattending treatment group students) were assigned to clusters on the basis of their sending school at the time of randomization in spring 2013. This resulted in all students in a summer site or classroom being defined to be in the same cluster, regardless of their actual regular-year schools, thus accounting for any clustering generated at the summer site and classroom levels.

The number of resulting clusters was less than 40 (because the number of summer sites was fewer than 40), so we were concerned about using the usual Huber-Eicker-White sandwich estimator cluster adjustment, which requires a large number of clusters to be valid. Instead, we used the bootstrap procedure discussed by Cameron, Gelbach, and Miller (2008) to calculate p-values for the null hypothesis that the treatment had zero effect.

Summary

In sum, we estimated four kinds of models:

1. random-effects, no baseline covariates
2. random-effects, baseline covariates
3. OLS with bootstrap-based inference, no baseline covariates
4. OLS with bootstrap-based inference, baseline covariates

Model 2 is our preferred specification, and the one for which we present results throughout *Ready for Fall? Near-Term Effects of Voluntary Summer Learning Programs on Low-Income Students' Learning Opportunities and Outcomes*. Model 4 was a robustness check that is a bit more conservative than Model 2, in that it can produce slightly larger standard errors. Our analyses confirmed that results from Models 2 and 4 were substantively the same.

Models 1 and 3 were checks that the inclusion of baseline controls was not driving the point estimate. We found that these point estimates were very similar to the preferred model because covariates were well balanced across the experimental groups due to randomization. The statistical significance of the estimates in Models 1 and 3 were not used to judge whether the summer programs had an effect.

Secondary analyses used simple extensions to Model 2. In the case of attendance and dosage models, the treatment assignment indicator was replaced with continuous or categorical variables for these mediators. When testing other mediators or moderators (such as student characteristics, or class/site characteristics), the variable of interest was interacted with the treatment indicator.

Analysis of Treatment Effect on the Treated

To estimate the effect of attendance in the summer program (i.e., the TOT), we had to account for the fact that selection into program attendance in the treatment group was endogenous. Therefore, we used randomization status as an instrumental variable for program attendance. Specifically, we estimated via two-stage least squares models of the form:

$$(2) \qquad Y_{ic} = \theta T_{ic} + d_c + X'_{ic}\beta + \varepsilon_{ic}$$
$$T_{ic} = \alpha Z_{ic} + d_c + X'_{ic}\varphi + \varepsilon_{ic}$$

where T_{ic} is a binary indicator of whether the student attended any days of the summer program. Now, the parameter of interest is θ, the coefficient on T_{ic} in the outcome equation. This parameter captured the average effect of the intervention for the subgroup of students who participated in the program (Angrist, Imbens, and Rubin, 1996; Bloom, 2006). P-values were derived using the wild cluster bootstrap-t procedure proposed by Cameron, Gelbach, and Miller (2008) for situations where the number of clusters is small. In this application, clusters were determined by summer site. For students in the control group, we used the summer site that most treated students in a student's regular-year school attended.

Multiple Hypotheses Testing

When performing multiple hypothesis tests on a data set, the chance of erroneously finding statistically significant results (i.e., Type I errors) increases as the number of tests increase. As a consequence, we adopted a standard corrective measure, which was to apply more stringent criteria for determining statistical significance. In practice, this translated into lowering the critical p-value for determining statistical significance to something less than 0.05. Exactly how much lower than 0.05 was determined by the number of tested hypotheses that pertain to a given domain of outcomes and the specific correction method used.

The exact methods and decisions to make when employing these corrections are a matter of debate among statisticians. After reviewing the spectrum of most conservative to most liberal options, we adopted a middle-ground position that adheres to the most detailed guidance that the Institute of Education Sciences has released on this topic (Schochet, 2008).

Consistent with WWC standards, we used the Benjamini-Hochberg method of controlling the false discovery rate (Benjamini and Hochberg, 1995). Following Schochet's guidance, we defined five student outcome domains pertinent to the summer learning demonstration:

1. mathematics outcomes
2. reading outcomes
3. social-emotional outcomes
4. school-year behavior (e.g., suspensions and expulsions)
5. school-year attendance

Only those hypotheses that belonged to a confirmatory category of an outcome domain were subject to corrections for multiple hypotheses testing within that domain. Exploratory hypotheses were not subject to multiple hypotheses corrections. For domain-specific confirmatory analyses in each district summer learning demonstration report, we adjusted downward the critical p-value for determining statistical significance according to the number of hypotheses tests belonging to that domain. Table B.1 enumerates the 15 confirmatory hypotheses tests conducted in each domain (mathematics, reading, and social-emotional).

Table B.1: Hypotheses Tested Within Each Outcome Domain for which Multiple Hypotheses Corrections Were Applied for Fall 2013 Results

Hypotheses Tested Using Near-Term Results	
1	Intent-to-treat (ITT) pooled estimate
2	Boston ITT estimate
3	Dallas ITT estimate
4	Duval ITT estimate
5	Pittsburgh ITT estimate
6	Rochester ITT estimate
7	ELL subgroup ITT estimate
8	FRPL subgroup ITT estimate
9	Low-achieving subgroup ITT estimate
10	Treatment-on-treated (TOT) pooled estimate
11	Boston TOT estimate
12	Dallas TOT estimate
13	Duval TOT estimate
14	Pittsburgh TOT estimate
15	Rochester TOT estimate

Appendix C

Data Collection

Academic Achievement

The primary outcomes of interest in the near-term analyses were students' performance on standardized assessments of their mathematics and reading achievement. We selected broad, general-knowledge, standardized assessments similar to state assessments and appropriate for the study population. The majority of students took the GMADE mathematics assessment and GRADE reading assessment by Pearson Education, which are 90-minute and 65-minute multiple-choice paper tests, respectively. These exams are offered at various levels that roughly correspond to grade levels, but are designed with flexibility to administer the test above or below the grade level indicated. (For example, Level 3 is nominally for third-graders, but is considered appropriate for second- or fourth-graders as well.) Students in the study were all fourth-graders in fall 2013 (with rare exceptions of grade retention or advancement). The project selected the Level 3 exam for the study students because they are generally low-performing students and because the tests would be administered very early in fourth grade.

There were two exceptions regarding which standardized assessments were administered to students in the study. The first occurred in Pittsburgh, where students took a Level 4 district-administered GRADE assessment (instead of Level 3 administered by the project) because the district was already administering this exam to all fourth-graders in fall 2014 as part of a district-wide initiative. The second occurred in Dallas, where students took the Texas spring 2013 assessment in Spanish rather than in English. For these students, the project administered the reading comprehension subtest of the Spanish-language Logramos assessment from Riverside Publishing instead of the GRADE.

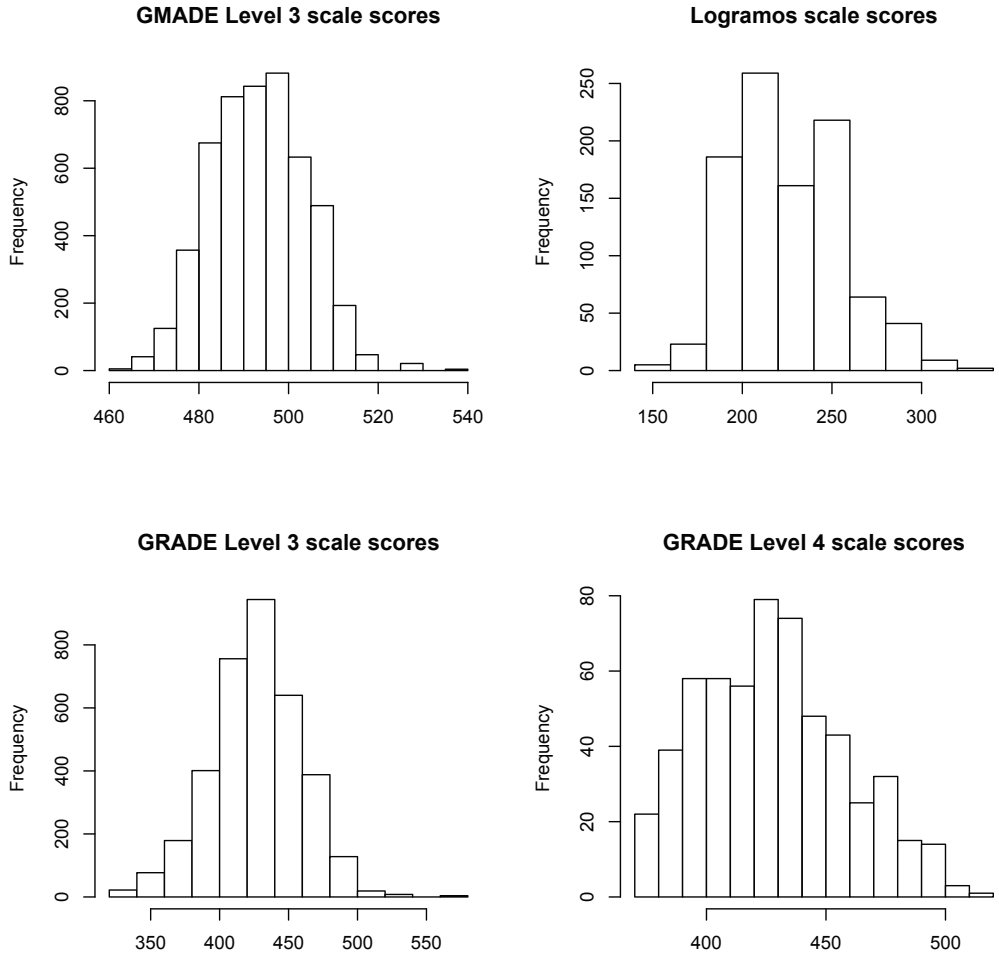
The project-administered assessments were given in fall 2013 during the third, fourth, and fifth weeks of the school year. The Wallace Foundation contracted the research firm Mathematica Policy Research to administer these assessments. Only participating students who were still enrolled in a public school within the five school districts were eligible for fall 2013 testing. Across all districts, the overall percentage of students who moved out of the study district or whose location was unknown at the time of testing was 7.4 percent. The highest percentage for any one district was 9.1 percent and the lowest for any one district was 4.8 percent. In total, we have mathematics scale scores for 5,127 students in the study (90.9 percent) and reading scale scores (either Logramos or GRADE) for 5,099 students in the study (90.4 percent). Descriptive information of the assessments' response rates is summarized in Table C.1.

Table C.1: Mathematics and Reading Assessments

	Mathematics			Reading		
District	**Type of Assessment Administered**	**Number of Students with Scorable Tests**	**Response Rate (%)**	**Type of Assessment Administered**	**Number of Students with Scorable Tests**	**Response Rate (%)**
Boston	Level 3 GMADE	870	90.9	Level 3 GRADE	874	91.3
Dallas	Level 3 GMADE	1,861	90.5	Level 3 GRADE administered by Mathematica. For ELL, Logramos reading comprehension subtest	1,857 (889 GRADE and 968 Logramos)	90.3
Duval	Level 3 GMADE	817	91.9	Level 3 GRADE administered by Mathematica	812	91.3
Pittsburgh	Level 3 GMADE	587	89.5	Level 4 GRADE administered by district	565	86.1
Rochester	Level 3 GMADE	992	91.9	Level 3 GRADE	991	91.8
Total		**5,127**	**90.9**		**5,099**	**90.4**

Figure C.1 score distributions on these assessments. There is no evidence of floor or ceiling effects, except for slight truncation of the distribution (floor effect) on the GRADE Level 4 exam that was administered in Pittsburgh.

Figure C.1: Distributions of Scores on Academic Assessments



Social-Emotional Outcomes

Broadly, social-emotional competence refers to the ability of students to successfully interact with other students and adults in a way that demonstrates an awareness of, and ability to manage, emotions in an age- and context-appropriate manner. To measure social and emotional well-being in the fall after the summer programs, RAND administered the Devereux Student Strengths Assessment—RAND Research Edition (DESSA-RRE) to school-year teachers who reported on the behaviors of individual study students. The DESSA is a strength-based assessment that assesses only positive behaviors rather than maladaptive ones.[10]

[10] Development of the DESSA was led by Paul LeBuffe, Valerie Shapiro and Jack Naglieri at the Devereux Center for Resilient Children and was made publicly available in 2009.

The DESSA-RRE comprises one scale of 27 items that RAND staff selected from the original item pool of 72 items on the DESSA (LeBuffe, Shapiro, and Naglieri, 2009). RAND selected these items based on their alignment with the school districts' stated goals for their summer programming. Drawing on student data from the DESSA national standardization sample, the developers determined that the pool of 27 items has a high degree of reliability; the 27 items and the corresponding coefficient Alpha are listed in Table C.2.

RAND administered the DESSA-RRE online to a study student's teacher of record beginning on the first day of the eleventh week of the school year. We selected this timing because it was the first time point at which the large majority of students would have been assigned to their teacher of record for at least four weeks. (Each district had determined the teacher of record no later than the first day of the tenth week of the school year.) In the DESSA-RRE, the rater is asked for each of 27 items to indicate on a 5-point scale how often the student engaged in each behavior over the past four weeks.

The survey took approximately five minutes to complete per student, and teachers were given a $20 Amazon gift card per survey completed. Teachers were required to answer 26 of 27 items for a survey to be deemed complete. We obtained responses from 84.0 percent of teachers of record and for 79.0 percent of the study students. Using district data for students who had left the district as of the time that we administered the DESSA, the effective response rate was 86.4 percent of the still-enrolled student sample.

With the fall 2014 DESSA-RRE results, we performed exploratory factor analyses and identified two subscales with high levels of internal consistency reliability. The items loading on each scale and the scales' reliability are also shown in Table C.2. We generated scores for the scales by averaging responses across the relevant items for each student. To these scales we assigned the names self-regulation (for items generally about students' ability to control their behavior and interactions) and self-motivation (for items generally about students' academic focus and drive).

Table C.2: DESSA-RRE Social-Emotional Behavior Scales

	Overall Social-Emotional Behavior Scale	Self-Regulation Scale	Self-Motivation Scale
Coefficient Alpha	*0.968*	*0.969*	*0.954*
During the past 4 weeks, how often did the child…			
1. Carry herself/himself with confidence	✓	✓	
2. Keep trying when unsuccessful	✓	✓	
3. Say good things about herself/himself	✓		✓
4. Compliment or congratulate someone	✓		✓
5. Show good judgment	✓		✓
6. Pay attention	✓		✓
7. Wait for her/his turn	✓		✓
8. Act comfortable in a new situation	✓		✓
9. Do things independently	✓		✓
10. Respect another person's opinion	✓		✓
11. Contribute to group efforts	✓		✓
12. Do routine tasks or chores without being reminded	✓	✓	
13. Perform the steps of a task in order	✓		
14. Show creativity in completing a task	✓	✓	
15. Share with others	✓	✓	
16. Accept another choice when his/her first choice was unavailable	✓		
17. Say good things about the future	✓	✓	
18. Stay calm when faced with a challenge	✓	✓	
19. Attract positive attention from adults	✓	✓	
20. Cooperate with peers or siblings	✓	✓	
21. Show care when doing a project or school work	✓	✓	
22. Make a suggestion or request in a polite way	✓	✓	
23. Learn from experience	✓	✓	
24. Work hard on projects	✓		✓
25. Follow rules	✓	✓	
26. Offer to help somebody	✓	✓	
27. Adjust well when going from one setting to another	✓	✓	

NOTE: The items were obtained from the DESSA and used with the permission of the Devereux Foundation.

Characteristics of Students in the Sample

Table C.3 shows the characteristics of students in the experiment according to whether they belong to the treatment or control group. Treatment and control group students differed in the

aggregate along some demographic characteristics because district demographics varied (for example, Dallas has a greater portion of Hispanic students and ELLs) and because the percentage assigned to treatment also varied by district. In Dallas, 50 percent of eligible applicants were randomized to the treatment group, whereas in each of the other four districts it was 60 percent. The combination of these two factors resulted in group differences on race/ethnicity, FRPL eligibility, and ELL variables. Once the varying proportion assigned to treatment was properly accounted for by controlling for strata, we did not see any statistically significant imbalance between the treatment and control groups on observed pretreatment characteristics listed in Table C.3.

Table C.3: Characteristics of Students in the Experiment

Combined Sample	Treatment Group Mean	Treatment Group SD	Control Group Mean	Control Group SD	Standardized Difference	p-value
Prior achievement						
Standardized spring 2013 mathematics score	0.017	0.918	0.003	0.933	0.015	0.570
Standardized spring 2013 language arts score	–0.013	0.913	–0.006	0.918	–0.008	0.768
Lower achieving	46.5%	0.499	46.2%	0.499	0.006	0.816
Spring 2013 mathematics score missing	8.3%	0.276	8.0%	0.271	0.012	0.661
Spring 2013 language arts score missing	9.4%	0.292	9.3%	0.291	0.001	0.968
Demographic characteristics						
ELL	29.3%	0.455	33.7%	0.473	–0.095	0.000
FRPL-eligible	86.2%	0.345	88.5%	0.319	–0.069	0.009
African American	49.2%	0.500	44.5%	0.497	0.094	0.000
Hispanic	38.0%	0.486	43.6%	0.496	–0.114	0.000
Other racial or ethnic category	12.7%	0.333	11.9%	0.323	0.027	0.317
Number	**3,194**	**—**	**2,445**	**—**	**—**	**—**

NOTE: SD = standard deviation

Student Survey Responses

To understand both the control and treatment groups' participation in any type of camp or summer school activity, the study included a short student survey about their activities during summer 2013. At the same time (fall 2013) that Mathematica Policy Research administered the GMADE, GRADE, and/or Logramos to students in the study, Mathematica Policy Research also administered a four-question survey to students. The survey was also translated into Spanish, and students who took the Logramos instead of the GRADE took the Spanish version of the student survey. Given the wide variety of summer programming available to students in the five school districts where the study occurred, the primary purpose of the survey was to gauge the contrast

between the treatment and control groups' exposure to any type of summer programming during summer 2013.

Table C.4 reports the number of students who completed the student survey and the proportions who answered each item. The table first summarizes the language in which students took the survey, and then the treatment and control group responses to each survey item. Below these items, we include composite variables that we constructed from the survey and that are used in the nonexperimental analyses that we reported in Chapter Three of *Ready for Fall? Near-Term Effects of Voluntary Summer Learning Programs on Low-Income Students' Learning Opportunities and Outcomes*. As the last row of Table C.4 indicates, almost one-third of the control group reported going to a summer camp or summer school for at least a few weeks in 2013.

Table C.4: Summer 2013 Student Survey Responses

	Treatment Group Respondents		Control Group Respondents	
	N	Percentage	N	Percentage
Language in which student took the survey				
English	2,429	82.9	1,737	78.8
Spanish	500	17.1	468	21.2
At home this last summer, I read a book or a magazine				
Never	370	12.7	321	14.7
A few times this summer	1,563	53.6	1,159	52.9
At least once a week	986	33.8	710	32.4
This last summer, I went to camp or summer school				
Did not go to camp or summer school	546	18.7	1,270	57.7
Went for a few days	238	8.1	125	5.7
Went for one week	154	5.3	111	5.0
Went for a few weeks	728	24.9	253	11.5
Went for at least a month	1,256	43.0	442	20.1
I did reading and writing at my camp or summer school this last summer				
Did not go to camp or summer school	544	18.8	1,269	58.6
No	168	5.8	230	10.6
Yes	2,179	75.4	665	30.7
I did mathematics at my camp or summer school this last summer				
Did not go to camp or summer school	545	18.8	1,272	58.7
No	185	6.4	363	16.7
Yes	2,169	74.8	533	24.6
Composite Survey Variables				
Attended camp with mathematics for at least “a few weeks”				
No	1,060	36.0	1,785	81.0
Yes	1,861	64.0	415	19.0
Attended camp with reading for at least “a few weeks”				
No	1,055	36.0	1,682	77.0
Yes	1,863	64.0	511	23.0
Attended camp for at least “a few weeks”				
No	938	32.0	1,506	68.0
Yes	1,984	68.0	695	32.0

NOTES: Not all students answered each of the four survey questions. Thus, the sum of respondents for each item is not always equal.

Summary of Teacher Survey Responses

RAND administered a five-to-ten–minute survey to all teachers of mathematics and language arts in summer 2013. The number of mathematics and language arts teacher respondents per district ranged from a minimum of 16 (out of 16 possible) to a maximum of 59 respondents (out of 66 possible) respondents per district. As shown in the second row of Table C.5, we obtained response rates of 89 to 100 percent of summer teachers in each district. We offered a $20 Amazon gift card to teachers who completed the survey. In all districts except for Boston, eligible teachers were first emailed a link to an electronic version of the survey. For nonresponders, RAND disseminated paper copies. In Boston, only paper was administered since Internet access was not available at each of the summer sites. The teacher survey was administered during the third through fifth week of the summer program within each district. The survey followed the same format in each district, with minor customization to reflect site names within each district and different dates and format of professional development offered to summer teachers. Following Table C.5 is an example of the Boston teacher survey.

Table C.5: Academic Teachers' Views of Their Summer Program, by District

Survey Item	District A	District B	District C	District D	District E
Number of Respondents	59	40	40	37	16
Participation rate (%)	89	89	100	95	100
Background					
Taught in summer program in 2012 (%)	25	35	40	32	63
Worked with same students during SY 2012–2013 (%)	23	38	46	46	69
Quality and structure of summer program					
This program is well managed and well organized. (%)	72	80	93	95	69
Support staff (e.g., camp counselors, paraprofessionals, instructional aides, tutors) provides necessary support in my classroom. (%)	81	88	90	86	94
There is a clear procedure for handling student discipline problems. (%)	88	83	80	81	88
The procedure for handling student discipline problems is effective. (%)	83	80	83	73	63
Climate and culture of summer program					
Administrators in this program care about students and teachers. (%)	97	100	100	95	100
Teachers listen to students when they have a problem. (%)	98	95	100	97	94
Faculty and staff make students feel cared for. (%)	100	98	100	100	94
Faculty and staff treat students with respect. (%)	100	98	100	100	94
Teachers enjoy teaching here. (%)	97	97	93	94	81
Faculty and staff remind students to be friendly and respectful to each other (%).	100	100	98	97	100
About students in summer program					
Due to student misbehavior, a great deal of learning time is wasted. (%)	34	18	38	57	56
Students enjoy this summer program. (%)	83	100	95	95	88
Students solve problems without fighting, or saying mean things. (%)	90	85	75	64	75
Students feel safe travelling to and being in this school. (%)	98	100	100	95	100
Students treat adults in school with respect. (%)	95	85	88	84	88
Children get into physical fights with other students at school at least once a week. (%)	7	5	8	35	38
Children are bullied and harassed by other students at least once a week. (%)	15	5	21	57	50

NOTES: Unless otherwise noted, values shown in table are the percent of respondents who agree or strongly agree.
Sources: RAND academic teacher survey data administered in summer 2013.

2013 Academic Teacher Survey (Boston Example)

Introduction

DEAR TEACHER,

The Wallace Foundation is funding your school district to study and augment its summer program. Your participation in this survey is important. Below are answers to some general questions you might have about the survey.

HOW LONG WILL THIS TAKE?
We estimate the survey will take about 5– 10 minutes to complete.

WHY SHOULD YOU PARTICIPATE IN THIS SURVEY?
Most importantly, your input will help influence how the program works next summer. ***You will also receive a $20 Amazon.com gift card as a token of appreciation for completing the survey.***

WHO IS BEING ASKED TO TAKE THIS SURVEY?
We are inviting all math and language arts teachers who work for the particular summer program that The Wallace Foundation is funding to complete this survey. Your participation is voluntary. However, we hope you will participate.

WHAT IS THE PURPOSE OF THIS SURVEY?
The purpose of this survey is to obtain information from teachers about your experience and advice on how your district could improve the summer program next year.

WHO IS CONDUCTING THIS SURVEY?
The RAND Corporation is conducting this survey on behalf of The Wallace Foundation.

WILL YOUR RESPONSES BE KEPT CONFIDENTIAL?
Yes. Your responses will not be shared with anyone working at your site this summer, at your school district, or anyone else outside the RAND research team. All responses that relate to or describe identifiable characteristics of individuals may be used only for statistical purposes and may not be disclosed, or used, in identifiable form for any other purpose, unless otherwise compelled by law.

HOW WILL YOUR INFORMATION BE REPORTED?
The information you provide will be combined with the information provided by others in statistical reports. No individually identifiable data will be included in the statistical reports.

If you have any questions about the study, please contact the Principal Investigator, Jennifer Sloan McCombs at 412-683-2300 x 5467 or by email at sloan@rand.org.

WE APPRECIATE YOUR TIME AND INPUT! WE HOPE YOU WILL ANSWER EVERY QUESTION.

About You and Your Students

1) Did you teach the following grade levels this summer?

*Mark only **one** response for each line.*

	Yes	No
	↓	↓
Third graders going into fourth grade	○	○
Fourth graders going into fifth grade	○	○

2) What is the name of the school or site where you work this summer?

*Mark only **one** response.*

- ☐ Summer Learning Project at the **Condon**
- ☐ Tenacity Summer Learning Project at the **Dever**
- ☐ Summer Learning Project at **Hale** Outdoor Learning Adventures
- ☐ **Jamaica Pond** Summer Learning Project
- ☐ **Boston Harbor** Summer Learning Project
- ☐ Summer Learning Project at the **Hennigan**
- ☐ Tenacity Summer Learning Project at the **Jackson-Mann**
- ☐ Tenacity Summer Learning Project at the **McKay**
- ☐ Summer Connections Program at **Thompson Island**
- ☐ Summer Learning Project at **Ponkapoag** Outdoor Center

About this Summer Program

3) To what extent do you agree or disagree with these statements about district professional development for this summer session?

*Choose only **one** for each line.*	**Agree a lot**	**Agree a little**	**Disagree a little**	**Disagree a lot**	**Did not attend**
The professional development:	↓	↓	↓	↓	↓
Prepared me to teach the **math curriculum** well.	○	○	○	○	○
Prepared me to teach the **language arts curriculum** well.	○	○	○	○	○
About **Voyager was worth my time** to attend.	○	○	○	○	○
About **American Reading Company was worth my time** to attend.	○	○	○	○	○
About **socio-emotional skill development was worth my time** to attend.	○	○	○	○	○

4) Did you teach math this summer?

Yes	No	
↓	↓	⇨ **If you did not teach math, skip to Q7!**
○	○	

5) Thinking about only your *math class* for students in your summer program, are the following statements true? *If you did NOT teach math this summer, please skip to question 7.*

*Mark only **one** response for each line.*	Yes	No
	↓	↓
I was **provided a written math curriculum**.	○	○
I was **provided a pacing guide** indicating which math topics are to be taught each week.	○	○
I received **lesson plans** to use for my math classes.	○	○
I **obtained the math instructional materials** (textbooks, curricular guides, lesson plans) with sufficient time to prepare for the first day of class.	○	○
I received information about students' **IEPs or special needs** prior to the first day of class.	○	○
I received **school-year data** on my students' prior math performance to help inform my instruction.	○	○
I received **summer math pretest results** for my students.	○	○

6) Thinking about only the *math curriculum* for students in your summer program, how much do you agree with the following statements? *If you did NOT teach math this summer, please skip to question 7.*

*Mark only **one** response for each line.*	**Agree a lot**	**Agree a little**	**Disagree a little**	**Disagree a lot**
	↓	↓	↓	↓
The **planned pacing** of the curriculum was **reasonable**.	○	○	○	○
Children's **math skills are improving** as a result of this program.	○	○	○	○
The math curriculum is **clear** for me to follow.	○	○	○	○
The math curriculum includes **fun, interesting activities** for children.	○	○	○	○
The math curriculum content is **too difficult** for a majority of students in my class.	○	○	○	○
The math curriculum content is **too easy** for a majority of students in my class.	○	○	○	○
The math curriculum **addresses gaps** that many students have **from last year**.	○	○	○	○

7) Did you teach language arts this summer?

Yes	No	
↓	↓	⇨ **If you did not teach language arts, skip to Q10!**
○	○	

8) Thinking about only the *English Language Arts class* for students in your summer program, are the following statements true? *If you did NOT teach English Language Arts this summer, please skip to question 10.*

*Mark only **one** response for each line.*	Yes ↓	No ↓
My school/site **grouped students by ability** in language arts for classroom assignments.	○	○
I was **provided a written language arts curriculum**.	○	○
I was **provided a pacing guide** indicating which language arts topics are to be taught each week.	○	○
I received **lesson plans** to use for my language arts classes.	○	○
I received information about students' **IEPs or special needs** prior to the first day of class.	○	○
I received school-year data on my **students' prior language arts performance** to help inform my instruction.	○	○
I received **summer language arts pretest data** for my students.	○	○

9) Thinking about only the *language arts curriculum* for students in your program this summer, how much do you agree with the following statements? *If you did NOT teach language arts this summer, please skip to question 10.*

*Mark only **one** response for each line.*	**Agree a lot** ↓	**Agree a little** ↓	**Disagree a little** ↓	**Disagree a lot** ↓
The **planned pacing** of the curriculum was **reasonable**.	○	○	○	○
Children's **language arts skills are improving** as a result of this program.	○	○	○	○
The language arts curriculum is **clear** for me to follow.	○	○	○	○
The language arts curriculum includes **fun, interesting activities** for children.	○	○	○	○
The language arts curriculum content is **too difficult** for a majority of students in my class.	○	○	○	○
The language arts curriculum content is **too easy** for a majority of students in my class.	○	○	○	○
The language arts curriculum **addresses gaps** that many students have **from last year**.	○	○	○	○
The language arts curriculum provides students **texts that are appropriate** for their reading level.	○	○	○	○

10) How much do you agree with these statements about the *quality and structure* of the summer program?

*Choose only **one** for each line.*	**Agree a lot** ↓	**Agree a little** ↓	**Disagree a little** ↓	**Disagree a lot** ↓
This program is **well managed** and well organized.	○	○	○	○
Support staff (e.g., camp counselors, paraprofessionals, instructional aides, tutors) provide necessary support in my classroom.	○	○	○	○
Administrators in this program **care** about students and teachers.	○	○	○	○
There is a **clear procedure** for handling student discipline problems.	○	○	○	○
The procedure for handling student discipline problems is **effective**.	○	○	○	○

11) How much do you agree with these statements about the *climate and culture* of the summer program?

*Choose only **one** for each line.*	**Agree a lot** ↓	**Agree a little** ↓	**Disagree a little** ↓	**Disagree a lot** ↓
Administrators in this program **care** about students and teachers.	○	○	○	○
Teachers **listen** to students when they have a problem.	○	○	○	○
Faculty and staff make students feel **cared for**.	○	○	○	○
Faculty and staff treat students with **respect**.	○	○	○	○
Teachers **enjoy** teaching here.	○	○	○	○
Faculty and staff remind students to be **friendly and respectful** to each other.	○	○	○	○

12) How much do you agree with these statements about the *students* in the summer program?

*Choose only **one** for each line.*	**Agree a lot** ↓	**Agree a little** ↓	**Disagree a little** ↓	**Disagree a lot** ↓
Due to student misbehavior, a great deal of **learning time is wasted.**	○	○	○	○
Students **enjoy** this summer program.	○	○	○	○
Students **solve problems** without fighting, or saying mean things.	○	○	○	○
Students feel **safe** travelling to and being in this school.	○	○	○	○
Students treat adults in school with **respect**.	○	○	○	○
Children get into **physical fights** with other students at school at least once a week.	○	○	○	○
Children are **bullied or harassed** by other students at least once a week.	○	○	○	○

Next Summer

13) If you could change only one thing about this summer program, what would it be?

14) If you could keep only one thing the same about this summer program, what would it be?

15) Do you have any other advice about how to improve this summer program next summer?

Background information

16) Did you teach in this program in summer 2012?

*Mark **one** response only.*

- ☐ Yes
- ☐ No

17) Have you worked with some of the students currently in this summer program in the prior school year (SY 2012–2013)?

*Mark **one** response only.*

- ☐ Yes
- ☐ No

You've completed the survey!

Thank you for your time and your valuable input. Your gift card will be sent to your email address.

If you would like to provide a secondary email address, please do so here:

_____

Classroom Observations

We conducted observations of academic and enrichment instruction in the five districts, using the same protocol for both. To create our observation protocol, we first reviewed some widely used validated instruments (such as The Classroom Assessment Scoring System measure developed at the University of Virginia and The Framework for Teaching developed by Charlotte Danielson).[11] These classroom observation instruments, however, were not necessarily designed to analyze aspects of the classroom that research about *summer programming* indicates are the most important features linked to improvements in student achievement. Consequently, RAND developed its own classroom observation protocol designed specifically to measure certain key aspects of our theoretical framework about how summer programs might lead to gains in student learning.

The classroom observation protocol was intended to gather information on the quality of instruction, time on academic task, and other aspects of the classroom, such as opportunities for social and emotional development.

Inter-Rater Agreement

We strove to ensure inter-rater agreement on the academic and enrichment instruction observation protocols. All observers across the five districts attended three days of training on how to use the instruments. At this training, observers watched and rated between eight and 12 videos per day of language arts, mathematics, and enrichment classrooms at elementary grade levels, completed the full observation protocols individually, and then assessed the degree of agreement on each item on the observation protocols to calibrate the observers' scoring of the classroom instruction. The group then extensively discussed rating disparities and recoded additional videos to further calibrate rating. Following the three-day training, four lead RAND researchers then established their own consistency in rating through pairwise correlations from ratings of additional classroom videos. The four lead researchers then participated in co-observations with the RAND staff responsible for field observations within each of the five school districts. They co-observed ten to 12 classroom segments (each of at least 15 minutes in duration) in the field during the first week of the summer program in each of the five districts. The lead researcher and the RAND co-observer collected their ratings on each of the items on the observation protocol and their ratings were compared across the ten to 12 classroom segments within each item.

[11] Classroom Assessment Scoring System accessible at http://www.teachstone.org/about-the-class/
The Framework for Teaching accessible at http://www.danielsongroup.org/

2013 Classroom Observations and Protocol (Dallas Example)

During summer 2013, RAND observed a total of 783 classroom sessions. Table C.6 displays the number of observations and subjects. For each classroom observed, the RAND staff person observed the entire class session from start to finish, coding a minute-by-minute time log during the class session, and then, at the end of class, answering 24 yes-no items plus six open-ended response items.

Table C.6: Number of Classroom Observations

Subject Area	Boston	Dallas	Duval	Pittsburgh	Rochester	Total
3go4 ELA	57	56	46	16	36	211
3go4 ELA—Writing	NA	NA	NA	NA	24	24
3go4 ELA—Walk to Intervention (ELA—WTI)	NA	NA	NA	NA	47	47
3go4 Enrichment	66	71	70	11	22	240
3go4 Math	34	56	38	16	34	178
3go4 Science	NA	NA	26	NA	NA	26
3go4 Success Maker	NA	NA	22	NA	NA	22
4go5 ELA	NA	4	11	4	NA	19
4go5 Math	NA	3	9	4	NA	16
Total	157	190	222	51	163	783

SOURCES: RAND classroom observations conducted in summer 2013.

The following fields composed the classroom observation template that RAND researchers used in Dallas. All sections of the observation protocol are uniform across the five school districts with the exception of the "fidelity to curriculum," which contains district-specific questions. The template was used within Excel, and observers filled out a new Excel file for each classroom subject session they observed.

Introduction

1. OBS. Observer initials.	
2. DATE. Date [MMDDYY]:	
3. SCH_ID. School/site Identifier [S1, S2, etc.]:	
4. TEACH_ID. Teacher Identifier [T1, T2, etc.]. If more than 1 lead teacher, string IDs together.	
5. SCHED_BEGIN. Class period scheduled beginning:	
6. BEGIN_REASON. Main reason, if any, for class starting at a different time:	
7. SCHED_END. Class period scheduled ending:	
8. END_REASON. Main reason, if any, for class ending at a different time:	
9. SCHED_MIN. Intended minutes of instructional period:	
10. ACTUAL_MIN. Actual minutes of the instructional period:	
11. OFF_TASK. % of class session that is off-task.	
12. ENACT_CLASS. Actual minutes as percentage of intended minutes.	
13. SUBJECT. Subject of class [M for math, language arts for reading, W for writing, ENR for enrichment, other]:	
14. STUD_BEGIN. Number of students (start):	
15. STUD_END. Number of students (end):	
16. STUD_AVGNUM. Average number of students in class period	

TIME LOG

17. Class segment activities

Directions: Characterize each class segment. Your time log should begin when a majority of students are in the room, regardless of whether the teacher has launched the lesson. The log should end when the majority of students leave the room. I = majority of students engaged in an instructional activity. NI = class is in session and a majority of students not engaged in an instructional activity (e.g., off-topic conversation, transition to next activity that lasts longer than 60 seconds, teacher involved in management activity and students not engaged in educational activity). NA = class not in session (bathroom break). You should watch and record the entire class period. ***Start a new row each time any 1 of the 4 class activities (I/NI/NA, DI, GP, IP) changes****. Segments are at least 60 seconds long.*

DI*: Teacher giving direct or explicit instruction about how to complete a task or explaining the academic content to complete the tasks. Teacher may walk through a few problems or examples. Might scaffold or do a think-aloud, explaining to students the strategy or skill required to complete the task. ("I do.") I-R-E can fall into this category.*

GP*: Teacher facilitates and students participate in instruction. This is where students apply the strategies that teacher demonstrated during the direct instruction. Teacher may ask a student to the board to complete a problem and then explain to the class what he/she did. Teacher may have all students fill out 1 problem and then group talk through that 1 problem before launching independent practice. ("We do.")*

IP*: Students have independent practice, whether in small groups or independent work. Student completes activities without consistent support from the teacher; e.g., reading a book and then filling out a worksheet. ("You do.")*

Time Begin:	Instruction (I), Noninstruction (NI), Class not in session (NA)	Majority of kids receiving direct instruction (DI)	Majority of kids engaged in guided practice (GP)	Majority of kids in independent practice (IP)	Summarize the major activity of the segment and positive or negative aspects of the segment
Actual class end time:					
Total min	*Total NI: _____* *Total I: ____* *Total NA: _____*	*Total DI:*	*Total GP:*	*Total IP :*	

FIDELITY TO THE CURRICULUM

18. Did the math lesson you observe have the following elements that every Voyager math lesson is supposed to contain? Skip if class observed was not mathematics.

Element	Description of element	Was element present (Y/N)?	Notes on positive and negative aspects of each element. Note "ok" if nothing notable
a. Getting Started (5 – 10 minutes)	Teacher should: • Review prerequisite skills • Model new concepts, skills, and strategies. • Emphasize math vocabulary • Engage students in understanding how math is used in real life.		
b. Guided Practice (10 minutes)	Teacher should: • Prompt students to verbally explain each math step, or explain if students are unable. Students should: • Verbalize each computational step.		
c. Independent Practice (15 minutes)	Students should: • Practice lesson content and previously learned skills on their own Teachers should: • Check student work and provide immediate correction/feedback.		
d. Test prep and error analysis (5 minutes)	Students should: • Practice math problem-solving with multiple choice and short answer formats Teachers should: • Assess daily progress.		

19. Did the language arts lesson you observe have the following elements that every National Geographic lesson is supposed to contain? Skip if class observed was not language arts.

Element	Description of element	Was element present (Y/N)?	Notes on positive and negative aspects of each element. Note "ok" if nothing notable.
a. Before Reading (30 minutes)	Teacher should: • Preview Concept Book • Read and discuss specific pages of text • Teach key concepts • Teach key vocabulary • Create some kind of graphic organizer (KWL chart, main idea, etc.) Students should: • Respond to questions • Contribute ideas • Complete learning masters		
b. During Reading (60 minutes)	Teacher should: • Mondays – Class read aloud • Teach specific comprehension strategy • Tuesdays – Thursdays - Read aloud/shared reading • Conduct small group instruction – Tiers 1, 2 & 3 • Use different leveled materials • Provide explicit instruction for Tiers 2 & 3 • Reinforce strategy instruction Students should: • Complete reading assignments • Complete learning masters • Engage in sustained silent reading or partner reading		
c. After reading (45 minutes)	Teacher should: • Read exemplar text • Provide mini-lesson on writing Students should: • Listen attentively to story and writing instruction • Draft, revise, and edit stories according to mini-lessons. • Publish and share stories		

20. Classroom practices that support student engagement. Indicate 1/0 for all items.

STATE GOAL		At the beginning of the lesson, the teacher stated or wrote what academic skills or strategies students will learn or practice during the lesson.
STATE PLAN		At the beginning of the lesson, the teacher clearly explained what students would do during the session.
WELL OILED		During most or all of the class period, a majority of students knew what they were supposed to do (or how to get the teacher's attention appropriately or ask for help). The class resembles a "well-oiled machine" where a majority of students know what is expected of them and how to go about doing it.
REDIRECT		When one or more students were off-task, the teacher noticed and effectively redirected that student or group to students to get back on task. If no students off-task, mark as 1.
ON-TASK		Are on-task. Students are focused, attentive, and not easily distracted from the task/project. They follow along with the staff and/or follow directions to carry on an individual or group task. Noise level and youth interactions can be high if youth are engaged in the expected task(s).
PRACTICE		All students got a chance to practice the skill/activity themselves during the class session.
PARTICIPATION		Encourage the participation of all. Regardless of gender, race, language ability, or other evident differences among youth, staff try to engage youth who appear isolated; they do not favor (or ignore) a particular youth or small cluster of youth. Staff need not force participation.
MONITOR ALL		During independent practice, the teacher monitors all, not just some, students as they work. (Check if the teacher consistently circulates through the space and looks at student work/activities while circulating.)
CHECK UNDERSTANDING		Teacher: (1) performs ongoing assessment during instruction by checking for understanding, and (2) addresses misunderstanding if and as they arise.

21. Evidence of other desirable practices. Indicate 1/0 for all items.

FRIENDLY		Are friendly and relaxed with one another. Youth socialize informally. They are relaxed in their interactions with each other. They appear to enjoy one another's company.
RESPECT		Respect one another. They refrain from derogatory comments or actions about an individual person and the work s/he is doing; if disagreements occur, they are handled constructively.
LIKE TEACHER		Show positive affect to staff. Youth interact with the staff, and these interactions are generally friendly interactions. For example, they may smile at staff, laugh with them, and/or share good-natured jokes.
COLLABORATE		Are collaborative. Youth work together/share materials to accomplish tasks.
LIKE STUDENTS		Show positive affect toward youth. Staff tone is caring and friendly; they use positive language, smile, laugh, or share good-natured jokes. They refrain from threats, cutting sarcasm, or harsh criticism. If no verbal interaction is necessary, staff demonstrate a positive and caring affect toward youth.
HIGHER STANDARD		Challenge youth to move beyond their current level of competency. Staff give constructive feedback that is designed to motivate youth, to set a higher standard, and meant to help youth gauge their progress. Staff help youth determine ways to push themselves intellectually, creatively, and/or physically.
ENTHUSIASM		All or almost all students exhibited obvious signs of enthusiasm for the class (e.g., jumping out of seat, eagerly raising hand, quickly and enthusiastically answering teacher's questions).
PERSIST		The teacher encouraged or supported students to persist at tasks that were difficult for them.
HELPFUL ADULTS		There was a helpful adult other than the teacher in the classroom. Helpful means the adult worked directly with students or the teacher to support students learning while they were in the room for a majority of the class time.
COMPLEX ACTIVITY		The activity is a complex one that requires multiple steps or progression of skills to complete well (e.g., sports, scavenger hunt, creating a map. *Not* Simon Says or free play).
EXPLICIT SOCIALSKILLS		The activity *explicitly* taught social skills such as cooperation with others, teaching of politeness or respectfulness, or required offers of help to others. Do not check if these skills were implicitly involved.
CHOICES		Within the structured activity, students were allowed to make individual choices (e.g., if working on an art project, students could choose what to draw or paint, or which details to include in a drawing or painting).
GROUP GOAL		Students' individual work contributes toward a group goal (e.g., individual students contributing to a class mural or book).

22. Evidence of undesirable practices. Indicate 1/0 for all items.

TOO HARD		The content was clearly too hard for a majority of the students.
TOO EASY		The content was clearly too easy for a majority of the students.
FACTUAL INACCURATE		The teacher provided or failed to correct factually inaccurate information that would confuse students about the content/skills they were to learn. (Do not count minor mistakes that do not relate to the skills being taught; e.g., stating today is Tuesday when it is Wednesday.)
UNCLEAR		The explanation of the instructional content was unclear or hard to follow. If you were a student in this class you would not know how to apply on your own the skill the teacher explained and/or demonstrated.
DISRESPECTFUL		In at least one instance, the teacher was disrespectful to students. This includes yelling at one or more students, using physical aggression, intentionally humiliating or ignoring a student, using discriminatory acts or derogatory language to students.
MISBEHAVIOR		There was one or more flagrant instance of student-to-student misbehavior. This includes a physical fight, persistent bullying, or persistent use of discriminatory or derogatory language.
UNSAFE		The learning/activity space was unsafe (e.g., broken glass on court, open chemical vats).
NO MATERIALS		A lack of materials impeded the teacher's ability to deliver the lesson or the students' ability to learn.
BEHAVIOR INTERRUPT		When the teacher addressed a student displaying an inappropriate behavior that derails the class (e.g. off-task, disruptive) the majority of the class was interrupted. The majority of students could not continue in their work/activity.
TEACHER DISENGAGED		The teacher responsible for the activity was disengaged in the classroom because of distractions by factors that were within her control (i.e., a teacher stopping by to have a conversation about the weekend, the teacher checking his/her cell phone, texting, or taking or making a personal call that was not related to an emergency, personal chat with co-teacher or paraprofessional while students are working).
ADULTS DISENGAGED		There were adults other than the teacher in the classroom who engaged in activities that distracted from learning (e.g. checking cell phone, interrupting the lesson, asking off-topic questions). Do not check if that distraction is isolated and brief. Also, do check if you know the person(s) is supposed to support instruction, such as a paraprofessional, but isn't for a majority of the class time. Don't check this item if an adult whose role you do not know is quietly observing a classroom.
BORED		All or almost all students in the class appeared bored throughout the class. Boredom characterized the class period.

Overall Reaction:

1) What did students learn from this lesson?
2) What, if any, were the main impediments or barriers to learning in this class?
3) Was the content of this class shallow, deep (higher order), or somewhere in between? Why?
4) Was the level of the teacher's questions shallow, deep, or somewhere in between? Why?
5) Did the teacher's knowledge of the content seem shallow, deep, or somewhere in between? Why?
6) Did the lower-performing students in the class receive support from the teacher or another adult in the class?

Student Attendance

Among the 3,194 students in the treatment group, 2,515 attended one or more days of the district 2013 summer program and 679 students (21 percent of the treatment group) never attended. Table C.7 displays attendance patterns across the districts. No-show rates ranged from a low of 8 percent to a high of 32 percent. However, among students who did attend at least one day, students attended in one district regularly, on average attending 83 percent of the available days. In two other districts, students attended, on average, 69 percent of the available days. We also observed variance in attendance across sites within districts.

Table C.7: Attendance Rates, by District

District	Number of Students in Treatment Group	No-Shows* (% of Students)	Average Attendance** (% of Days Attended)	Attendance Range by Site (% of Days Attended)
District A	1,029	27	70	61–80
District B	534	32	83	71–88
District C	574	17	80	70–92
District D	647	8	69	68–72
District E	410	19	69	65–71
Total	**3,194**	**21**	**74**	

* No-shows are the percentage of students who did not show up for a single day of the program.
** Average attendance is percentage of days attended by students who attended one or more days.
SOURCES: District summer 2013 attendance data.

As presented in Table C.8, among treatment students, about half attended consistently (85 percent or more of the days) while about 7 percent attended only during the first week of the program and never returned.

Table C.8: Attendance Patterns, by District (by Percentage)

District	Average Daily Attendance	Attended 85% or More of Summer Program Days	Attended 5 or Fewer Days of the Summer Program	Attended 1 or More Days of the First Week of the Program but Never Thereafter
District A	70	52	19	13
District B	83	72	3	2
District C	80	60	6	4
District D	69	42	13	3
District E	69	46	13	7
Total	74	54	12	7

SOURCES: District summer 2013 attendance data.

Across the districts, as shown in Table C.9, among the 2,445 students in the control group, 114 (5 percent of the control group) were accidentally allowed to attend one or more days of the summer program. In ITT analyses, which compare outcomes for all treatment students against the outcomes of all control students regardless of whether they ever attended the program, no-shows and crossovers can result in underestimation of the effect of attending the summer program.

Table C.9: Noncompliance with Experimental Assignment

	Treatment Group (N=3,194)		Control Group (N=2,445)		
District	Number of No-Shows*	Percentage of No-Shows	Number of Crossovers**	Percentage of Crossovers	Average Percentage of Days Attended by Crossovers
District A	280	27	109	11	81
District B	173	32	4	1	43
District C	96	17	0	0	0
District D	52	8	0	0	0
District E	78	19	1	0	100
Total	**679**	**21**	**114**	**5**	**80**

* No-shows are enrolled students who did not attend any days of the summer program.
** Crossovers are students in the control group who were accidentally allowed to attend one or more days of the summer program.

Appendix D

Hypothesized Mediators and Moderators of Summer Program Effects

Attendance and Dosage: Amount of Instructional Time Received

Our primary hypothesis is that the amount of instructional time a student receives mediates (is part of the causal pathway for) the effect of the summer program on that student's outcomes. Therefore a primary objective of the analysis is to estimate as accurately as possible the amount of instruction students received. To increase the accuracy of attendance data, we collected and audited each week of the five- to six-week summer program student-level daily attendance data from each of the five school districts. To observe instructional time students received within the program day, RAND staff followed each student classroom cohort for at least one entire day during the summer sessions. For example, if a site enrolled three classrooms of students—"Group A," "Group B," and "Group C"—RAND visited the site for three full days in summer 2013 to follow each classroom cohort for a school day. Thus, RAND observed each student cohort's mathematics and language arts class at least once, and a majority of their enrichment programming courses[12] at least once during the program.

As we observed classrooms, we kept a time log recording when classes were scheduled to begin and end, the minute the majority of students were in the room and the teacher launched or ended the session, and minute-by-minute notes on class segments to track instructional and noninstructional time during the enacted class period (see 2013 Classroom Observation Protocol in Appendix C). For example, we recorded that a class was scheduled to begin at 10 a.m., actually launched at 10:11 a.m., lost a combined total of six minutes to noninstructional activities such as a bathroom break, ended at 10:59 a.m., and was scheduled to end at 11:00 a.m.

With these classroom observation data linked to student classroom rosters, we created student-level mathematics/language arts dosage indicators that equal the product of the following three measures: (1) the number of days a given student attended the summer program, multiplied by (2) the average number of hours that observed mathematics/language arts classes lasted (meaning the enacted time from class launch to class wrap-up, regardless of scheduled class time), averaged across the subject-relevant classes RAND observed within a given site in summer 2013, multiplied by (3) the

[12] Due to both the simultaneous enrichment activity rotations at some sites and the conducting of teacher interviews during some enrichment sessions in the second half of the summer session, RAND did not observe all classroom cohort-enrichment activity combinations at least once.

average percentage of enacted class time that was devoted to instruction. Table D.1 summarizes the average and the range of the number of instructional hours that treatment group students (including no-shows) received in mathematics and in language arts by district.

To increase the reliability of (2) and (3), we first averaged to the classroom (when classrooms were observed more than once), then to teachers (when teachers were observed more than once), and then to site.

Table D.1: Distribution of Instructional Hours that Treatment Group Students Received

	Language Arts			Mathematics		
District	Average Number of Instructional Hours	Minimum Number of Instructional Hours	Maximum Number of Instructional Hours	Average Number of Instructional Hours	Minimum Number of Instructional Hours	Maximum Number of Instructional Hours
District A	23.6	0	52.7	18.2	0	42.1
District B	24.3	0	55.2	21.4	0	44.1
District C	30.1	0	60.2	14.9	0	32.8
District D	21.4	0	34.3	15.8	0	25.1
District E	13.3	0	27.5	15.3	0	34.6
Total	**23.0**	**0.0**	**60.2**	**17.2**	**0.0**	**44.1**

Once we developed an estimate of the number of instructional hours an attendee received during summer 2013, we then performed a Generalized Additive Mixed Models (GAMM) analysis to identify discrete changes, or cut points, in the relationship between hours and mathematics/reading achievement. In contrast to linear models, GAMM enables us to examine nonlinear relationships between the two variables. For example, if 15 or fewer instructional hours are not associated with any increase, but there are incremental increases beyond those hours, then the GAMM curve would be relatively flat until 16 hours, with an increasing slope thereafter. The locations and number of inflection points in the GAMM curves are a matter of interpretation. We engaged a statistician who was not directly involved in this project to examine the curves and identify the inflection points, which we then used to create bins, or levels of dosage. Because the formation of these bins was done after examining the relationship between dosage and outcomes, they enabled us to specify levels of dosage for our outcomes models that were likely to be sensitive to the different dosage levels' effects on achievement. However, this *a priori* inspection of the relationship to outcomes is only appropriate for noncausal interpretations.

The same procedure was also applied to attendance. Because students self-select into their level of participation in summer programming, the GAMM analyses represent an exploratory technique that examines the relationship between intensity of treatment and a given outcome, as well as any nonlinearities that may exist in that relationship.

Table D.2 displays the resulting ranges for levels of dosage and attendance.

Table D.2: Thresholds for Dosage and Attendance Categories

	Mathematics		Reading		Social-Emotional
	Dosage	Attendance	Dosage	Attendance	Attendance
Low	< 13	< 22	< 31	< 22	< 18
Medium	≥13 to < 26		≥31 to < 39		
High	≥26	≥22	≥39	≥22	≥18

Creation of Relative Opportunity for Individual Attention

In our recent review of Kim and Quinn's (2013) meta-analysis, we noted their observation that programs that had small class size and high dosage appeared to be associated with positive outcomes. To explore whether this relationship held for the Wallace summer study, we used our data on dosage and class size to create a synthetic variable we term students' "relative opportunity for individual attention." We hypothesized that student outcomes would be mediated by not only the total amount of instructional time received in the summer session, but also by the number of students over which a lead teacher's attention was spread. In other words, we hypothesized that smaller class sizes would enhance the effective "dose" of instructional time received by a focal student as compared to another student with the same amount of instructional hours but within a larger class. Consequently, we developed a measure of dosage-by-class size, which was simply the division of a student's mathematics/language arts instructional hours by that student's average mathematics/language arts class size. "Relative opportunity for individual attention" is intended as a proxy for, rather than a direct measure of, individualized attention. We interpreted this variable as the relative amount of instructional attention from a teacher that theoretically could have been available to each student over the entire summer.

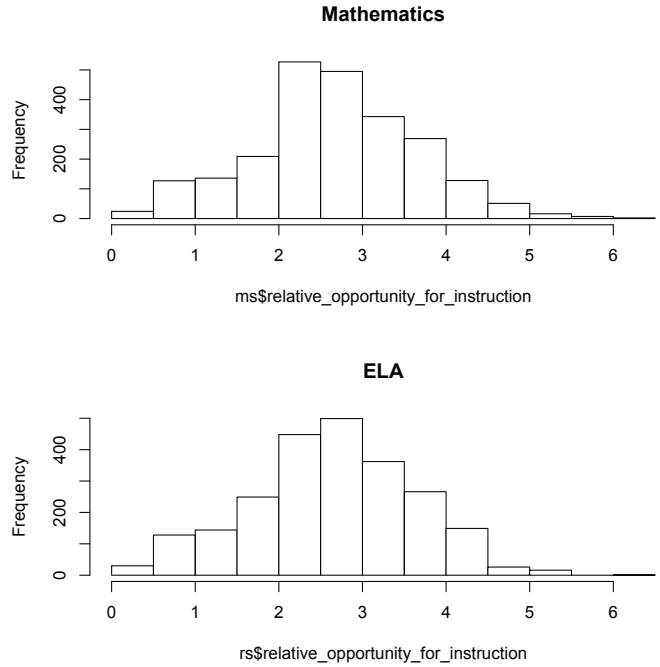
To create this measure, we first calculated the average number of students present in each language arts/mathematics class (shown in Table D.3) by applying districts' student-level attendance data to summer 2013 language arts/mathematics classroom rosters. Each student in the treatment group who attended one or more days of the summer program was associated with his assigned language arts/mathematics classroom size.

Table D.3: Distribution of Average Language Arts/Mathematics Class Sizes

	Language Arts			Mathematics		
District	**Average Class Size**	**Minimum Class Size**	**Maximum Class Size**	**Average Class Size**	**Minimum Class Size**	**Maximum Class Size**
District A	9.3	4.2	14.7	9.3	4.2	14.7
District B	8	3	13.2	8	3	13.2
District C	11.3	6.9	17.1	11.3	6.9	17.1
District D	12.2	7.9	16.9	12.2	7.9	16.9
District E	14.4	9.9	17.1	14.4	9.9	17.1
Overall	**10.4**	**3.0**	**17.1**	**10.4**	**3.0**	**17.1**

After dividing a focal student's total instructional hours by his or her average class size, we then applied a mathematical transformation (square root) to obtain a more normal distribution, and then normalized the values to have a standard deviation of one. Finally, we assigned a value of zero for this "relative opportunity for individual attention" measure to students in the treatment group who never attended the summer 2013 program. Figure D.1 shows the distributions of this variable for mathematics and language arts students (excluding no-shows).

Figure D.1: Distributions of the Relative Opportunity for Individual Attention Variable



Scales Created from Teacher Survey and Classroom Observation Data

We further hypothesized that, in addition to the amount of instructional time students received, the academic quality, the appropriateness of the mathematics/language arts curriculum, the match between student grade level and a teacher's prior year grade level assignment, the opportunity for social-emotional development, and the general degree of safety and order within summer sites would moderate the effect of summer programming on attendees. To test these hypotheses, we identified individual items or sets of items from classroom observation and/or teacher survey data to serve as measures for these constructs.

We generated six total scales by first averaging item-level responses for each scale across the relevant classroom or survey items for each classroom observation or respondent. In addition to these six scales, we also tested whether teaching the sending or receiving grade level of students moderated the summer programming effects on students. Summary statistics for these scales are shown in Table D.4.

Table D.4: Summary Statistics of Hypothesized Mediators

Mediator	Level at Which Mediator Measured (data source)	N	Mean	Minimum	Maximum
Quality of language arts instruction scale	Classroom (RAND classroom observations)	199	5.0	2.7	7.1
Quality of mathematics instruction scale	Classroom (RAND classroom observations)	195	5.2	2.6	7.0
Opportunity for social-emotional development scale	Site (RAND classroom observations)	37	4.6	3.2	6.2
Appropriateness of language arts curriculum	Classroom (teacher survey)	139	16.0	8.0	20.0
Appropriateness of mathematics curriculum	Classroom (teacher survey)	145	13.0	5.0	16.0
Student discipline and order scale	Site (teacher survey)	37	16.7	11.4	20.0
Student's language arts teacher taught third or fourth grade in SY 2012–2013	Classroom (staff rosters)	155	0.56	0	1
Student's mathematics teacher taught third or fourth grade in SY 2012–2013	Classroom (staff rosters)	153	0.66	0	1

The survey or classroom observation items that were included in each scale and the estimates of internal consistency reliability (coefficient Alpha) are provided in Tables D.5 and D.6. Note that some of the scales shown in these tables had coefficient Alpha values of less than 0.70, meaning that high levels of measurement error in the scales might have influenced the findings reported. However, this error would result in overly conservative, rather than upwardly biased, estimates of the moderating effect of these constructs.

Quality of Instruction

This scale is derived from RAND summer 2013 classroom observation data and calculated at the student's language arts/mathematics classroom level by first summing the items in the scale (listed in Table D.5). For mathematics and language arts instructional quality measures, we wanted to retain classroom-level measures even though measures derived from observations are prone to large error due to the small number of observations.

To explore improving the accuracy of these measures, we applied a simplified version of small-area estimation shrinkage (McCaffrey, Han, and Lockwood, 2013). We did this by fitting a hierarchical linear model with the instructional quality scale as the dependent variable, fixed effects for each district, and random effects for classes nested in teachers, who were, in turn, nested in summer sites. From these models, we then derived predicted, shrunken estimates for instructional quality for each class. Compared to the raw estimates, the shrunken estimates exhibited improved distributions and correlations with outcomes. However, both the raw and shrunken estimates of instructional quality produce the same inferences when tested for relationships with treatment effects. We plan to continue to explore ways to deal with the measurement error in these variables.

Table D.5: Mathematics/Language Arts Instructional Quality Items and Internal Consistency Reliability Estimates

Scale Items
Coefficient Alpha for Mathematics: 0.632
Coefficient Alpha for Language Arts: 0.672
1. Range from 0–1 point = observed % of mathematics/language arts class time that was spent on instruction. Scaled so 0 = min observed percent on instruction, and 1 = max observed % on task.
2. 1 point if "The teacher exhibited obvious signs of enthusiasm about the content of the class."
3. 1 point if "Large majority of students are on-task throughout the class. Students are focused and attentive to the task/project."
4. 1 point if there were no incidences of "The teacher provided or failed to correct factually inaccurate information that would confuse students about the content/skills they were to learn."
5. 1 point if "Teacher explained purpose of class in terms of real-world relevance."
6. 1 point if there were no incidences of "Teacher's explanation of the instructional content was unclear or hard to follow."
7. 1 point if "Teacher: (1) performs ongoing assessment throughout the whole class period by checking for students' understanding of content, and (2): addresses misunderstanding if and as they arise.
8. 1 point if rated no: "When the teacher disciplined students, the majority of the class was interrupted for a long period."
9. 1 point if rated no: "The teacher responsible for the activity was disengaged in the classroom because of distractions by factors that were within her control."
10. 1 point if yes: "All or almost all students exhibited obvious signs of enthusiasm for the class throughout the class period (e.g., jumping out of seat, quickly and enthusiastically answering teacher's questions)."

Appropriateness of the (Mathematics/Language Arts) Curriculum

We hypothesized that the curriculum that teachers deem appropriate for their students—which we define as a combination of perceptions about reasonable pacing, clarity of curriculum, addressing the right gaps in student knowledge and skills, and being fun for students—would enhance the effectiveness of summer programming in boosting student achievement. This scale is thus derived from the academic teacher survey. It is a teacher-level construct and associated with the treatment students assigned to that mathematics/language arts teacher. In the survey, teachers who reported teaching mathematics during summer 2013 were prompted to answer mathematics curriculum questions, with a parallel structure for language arts teachers. Teachers who taught both subjects were asked to complete both sets of curriculum questions. The mathematics curriculum scale includes four items and the language arts curriculum scale includes five.

Table D.6: Appropriateness of Mathematics/Language Arts Curriculum Scale Items and Internal Consistency Reliability Estimates

Scale Items
Coefficient Alpha for Language Arts: 0.749
Coefficient Alpha for Mathematics: 0.741
1-4 points on Likert scale. The planned pacing of the curriculum was reasonable.
1-4 points on Likert scale. The mathematics curriculum is clear for me to follow.
1-4 points on Likert scale. The mathematics curriculum addresses gaps that many students have from last year.
1-4 points on Likert scale. The mathematics curriculum includes fun, interesting activities for children.
1-4 points on Likert scale: [for Language Arts only]: The language arts curriculum provides students texts that are appropriate for their reading level.

Site Discipline and Order Scale

Like the social-emotional development scale, this is a site-level scale, but derived from teacher survey data within sites (see Table D.7). The working hypothesis here was that attendance at sites that teachers deemed safe (free of bullying and fighting) and that teachers deemed to have a clear set of procedures for discipline would enhance the effect of summer programming on student achievement. Items in the scale were first summed within a respondent, and then an unweighted average of respondents was taken to develop a site-level scale score.

Table D.7: Site Discipline and Order Scale Items, and Internal Consistency Reliability Estimates

Scale Items
Coefficient Alpha for Language Arts: 0.811
1. 1–4 points on Likert scale. “Children are bullied and harassed by other students at least once a week.”
2. 1–4 points on Likert scale. “Children get into physical fights with other students at school at least once a week.”
3. 1–4 points on Likert scale. “The procedure for handling student discipline problems is effective.”
4. 1–4 points on Likert scale. “There is a clear procedure for handling student discipline problems.”
5. 1–4 points on Likert scale. Reverse coded. “Due to student misbehavior, a great deal of learning time is wasted.”

Stand-Alone Moderators

Finally, we hypothesized that having a teacher who worked in a proximate grade level during the prior school year (either third or fourth grade) would moderate the effect of summer programming on student mathematics and reading achievement. This was because those teachers would theoretically be versed in the school-year academic standards that applied either in the year preceding or the year following the students’ summer session, and they would be familiar with the most common gaps between third- and fourth-graders’ knowledge and these standards. For this item, we simply associated the dichotomous indicator with each treatment group attendee via classroom rosters.

Appendix E

Results from Regression Models with Covariates

In this appendix, we report in tabular format the results narratively described in Chapter Five of *Ready for Fall? Near-Term Effects of Voluntary Summer Learning Programs on Low-Income Students' Learning Opportunities and Outcomes.*

Table E.1: Intent-to-Treat Results, Overall and by District

	Mathematics (Control = 2,205 Treat = 2,921)	Reading (Control = 2,196 Treat = 2,902)	Social-Emotional (Control = 1,903 Treat = 2,542)
	Estimate (Std. Error)	Estimate (Std. Error)	Estimate (Std. Error)
Overall	0.09* (0.02)	0.01 (0.02)	0.01 (0.03)
District A	0.08* (0.03)	0.06 (0.03)	0.04 (0.04)
District B	0.11 (0.05)	–0.03 (0.04)	0.00 (0.06)
District C	0.09 (0.04)	0.04 (0.04)	0.10 (0.06)
District D	0.06 (0.05)	–0.05 (0.03)	–0.04 (0.05)
District E	0.13* (0.05)	–0.02 (0.04)	–0.10 (0.08)

NOTE: * indicates statistical significance at the p<0.05 level after applying the Benjamini-Hochberg correction for multiple hypothesis tests.

Table E.2: Counts of Students Participating in Subgroup Analyses

		Mathematics		Reading		Social-Emotional	
		N Control	N Treat	N Control	N Treat	N Control	N Treat
ELL	no	1,447	2,057	1,435	2,045	1,268	1,810
	yes	758	864	761	857	635	732
FRPL-eligible (excl. Boston)	no	182	258	179	255	162	233
	yes	1,677	2,139	1,670	2,120	1,466	1,898
Below median on prior achievement	no	1,100	1,460	1,092	1,459	946	1,256
	yes	1,105	1,461	1,104	1,443	957	1,286

NOTE: To calculate students' prior achievement, we used the control group's performance on the common GRADE/GMADE posttests to scale each district's pretests. The posttests were scaled to have a mean of zero, a standard deviation of one, then we calculated the mean and standard deviation of the pretests for the control group subsample in each district, excluding crossovers. Then, each district's pretests for the whole sample were scaled to have the district's calculated means and standard deviations.

Table E.3: Results of Subgroup Analyses

	Mathematics	Reading	Social-Emotional
ELL	0.02	–0.01	0.03
	(0.04)	(0.04)	(0.06)
FRPL-eligible (excl. Boston)	0.04	0.05	0.03
	(0.06)	(0.05)	(0.09)
Below median on prior achievement	–0.06	0.04	0.09
	(0.03)	(0.03)	(0.05)

NOTE: Standard error is shown in parenthesis below each effect estimate.

Table E.4: Results of Mathematics Subscale Analyses

	N Control	N Treat	Estimate (Std. Error)
Concepts and communication	2,208	2,925	0.09*** (0.02)
Operations and computation	2,207	2,921	0.11*** (0.03)
Process and application	2,206	2,918	0.04* (0.02)

NOTE: *** p<0.001, * p<0.05; p-values are not adjusted for multiple hypothesis tests. Standard error is shown in parenthesis below each effect estimate.

Table E.5: Results of Reading Subscale Analyses

	N Control	N Treat	Estimate (Std. Error)
Comprehension (GRADE and Logramos)	1,953	2,504	0.01 (0.02)
Vocabulary (GRADE only)	1,460	1,971	0.01 (0.02)
Listening comprehension / oral language (GRADE only)	1,522	2,044	–0.04 (0.03)

NOTE: Analysis excludes Pittsburgh because the district administered Level 4 of the GRADE and because their scanned score results were largely missing subscale scores. Standard error is shown in parenthesis below each effect estimate.

Table E.6: Results of Social-Emotional Subscale Analyses

	N Control	N Treat	Estimate (Std. Error)
Self-regulation	1,903	2,542	0.01 (0.03)
Self-motivation	1,903	2,542	0.02 (0.02)

NOTE: Standard error is shown in parenthesis below each effect estimate.

Table E.7: Results of Treatment-on-the-Treated Analyses

	Mathematics	Reading	Social-Emotional
Overall	0.11* (0.03)	0.02 (0.02)	0.01 (0.02)
District A	0.10 (0.07)	0.09 (0.05)	0.05 (0.04)
District B	0.16 (0.08)	–0.04 (0.05)	0.00 (0.06)
District C	0.11 (0.07)	0.05 (0.03)	0.12 (0.05)
District D	0.06 (0.04)	–0.05 (0.03)	–0.03 (0.03)
District E	0.14 (0.11)	0.01 (0.03)	–0.09 (0.10)

NOTE: * indicates statistical significance at the p<0.05 level after applying the Benjamini-Hochberg correction for multiple hypothesis tests. In this analysis, the treated are defined as students assigned to the treatment group who attended at least one day of their district's summer program.

Table E.8: Nonexperimental Linear Effect of Attendance and Dosage

	Mathematics	Reading	Social-Emotional
Attendance	0.06*** (0.02)	0.02 (0.02)	0.06* (0.03)
Dosage (instructional hours)	0.05*** (0.01)	0.02 (0.01)	

NOTES: *** p<0.001, * p<0.05; p-values are not adjusted for multiple hypothesis tests. Standard error is shown in parenthesis below each effect estimate.

Table E.9a: Nonexperimental Effect of Attendance Categories

	Mathematics		Reading		Social-Emotional	
	N	Estimate (Std. Error)	N	Estimate (Std. Error)	N	Estimate (Std. Error)
No-show	588	–0.01 (0.03)	584	–0.03 (0.03)	494	0.03 (0.04)
Low	1,054	0.07** (0.02)	1,057	–0.01 (0.02)	621	–0.06 (0.04)
High	1,279	0.14*** (0.02)	1,261	0.04 (0.02)	1,427	0.04 (0.03)

Note: *** p<0.001, ** p<0.01; p-values are not adjusted for multiple hypothesis tests.

Table E.9b: Nonexperimental Effect of Dosage Categories

	Mathematics		Reading	
	N	Estimate (Std. Error)	N	Estimate (Std. Error)
No-show	588	–0.01 (0.03)	584	–0.03 (0.03)
Low	493	0.05 (0.03)	1152	–0.01 (0.02)
Medium	1,011	0.10*** (0.02)	493	0.03 (0.03)
High	829	0.16*** (0.03)	673	0.04 (0.03)

NOTE: *** p<0.001; p-values are not adjusted for multiple hypothesis tests.

Table E.10: Nonexperimental Effects of Camp Attendance According to Student Survey

	Mathematics		Reading		Social-Emotional	
	N	Estimate (Std. Error)	N	Estimate (Std. Error)	N	Estimate (Std. Error)
Attended camp w/mathematics >= a few weeks, control	413	0.02 (0.03)				
Did not attend camp w/mathematics ≥ a few weeks, treatment	1,049	–0.01 (0.02)				
Attended camp w/mathematics ≥ a few weeks, treatment	1,849	0.15*** (0.02)				
Attended camp w/reading ≥ a few weeks, control			504	0.01 (0.03)		
Did not attend camp w/reading ≥ a few weeks, treatment			1,035	–0.04 (0.02)		
Attended camp w/reading ≥ a few weeks, treatment			1,847	0.04* (0.02)		
Attended camp ≥ a few weeks, control					576	–0.04 (0.04)
Did not attend camp ≥a few weeks, treatment					778	0.01 (0.04)
Attended camp ≥ a few weeks, treatment					1,691	0.00 (0.02)

Note: *** p<0.001, * p<0.05; p-values are not adjusted for multiple hypothesis tests. Reference group is control group students reporting they did not attend camp (with mathematics or reading) for at least a few weeks.

Table E.11: Nonexperimental Estimates of the Effect of Relative Opportunity for Instruction

	Mathematics	Reading
Relative opportunity for instruction	0.01	0.00
	(0.02)	(0.02)

NOTE: Standard error is shown in parenthesis below each effect estimate.

Table E.12: Nonexperimental Estimates of Moderation

	Mathematics	Reading	Social-Emotional
Mathematics classroom instructional quality scale	0.00 (0.01)		
Language arts classroom instructional quality scale		0.04** (0.01)	
Site climate scale	0.01 (0.02)	–0.01 (0.02)	0.01 (0.03)
Appropriateness of mathematics curriculum	0.00 (0.02)		
Appropriateness of language arts curriculum		0.00 (0.01)	
Student discipline and order	0.00 (0.02)	0.03* (0.02)	0.04 (0.03)
Mathematics teacher taught 3rd or 4th grade	0.01 (0.03)		
Language Arts teacher taught 3rd or 4th grade		0.07* (0.03)	

NOTE: ** p<0.01, * p<0.05; p-values are not adjusted for multiple hypothesis tests. Standard error is shown in parenthesis below each effect estimate.

References

Angrist, Joshua D., Guido W. Imbens, and Donald B. Rubin, Identification of Causal Effects Using Instrumental Variables, *Journal of the American Statistical Association,* Vol. 91, No. 434, 1996, pp. 444–455.

Benjamini, Yoav, and Yosef Hochberg, "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing," *Journal of the Royal Statistical Society, Series B (Methodological)*, Vol. 57, No. 1, 1995, pp. 289–300.

Bloom, Howard S., *The Core Analytics of Randomized Experiments for Social Research MDRC Working Papers on Research Methodology*, MDRC Working Papers on Research Methodology, New York: MDRC, 2006. As of October 22, 2014: http://www.mdrc.org/sites/default/files/full_533.pdf

Cameron, A. Colin, Jonah B. Gelbach, and Douglas L. Miller, "Bootstrap-Based Improvements for Inference with Clustered Errors," *Review of Economics and Statistics*, Vol. 90, No. 3, 2008, pp. 414–427.

Duflo, Esther, Rachel Glennerster, and Michael Kremer, "Using Randomization in Development Economics: A Toolkit," in T. P. Schults and J. A. Strauss, eds., *Handbook of Development Economics*, Vol. 4, Amsterdam: North-Holland, Elsevier B.V., 2007, pp. 3862–3895.

Eicker, Friedhelm, "Limit Theorems for Regression with Unequal and Dependent Errors," *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, Vol. 1, Berkeley, Calif.: University of California Press, 1967, pp. 59–82.

Huber, Peter J., "The Behavior of Maximum Likelihood Estimates Under Nonstandard Conditions," *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, Vol. 1, Berkeley, Calif.: University of California Press, 1967, pp. 221–233.

Imbens, Guido W., *Experimental Design for Unit and Cluster Randomized Trials*, Cambridge, Mass.: Harvard University, Department of Economics, 2011.

Kim, James S., and David M. Quinn, "The Effects of Summer Reading on Low-Income Children's Literacy Achievement from Kindergarten to Grade 8: A Meta-Analysis of Classroom and Home Interventions," *Review of Educational Research,* Vol. 83, No. 3, 2013, pp. 386–431.

LeBuffe, Paul, Valerie Shapiro, and Jack Naglieri, *Devereux Student Strengths Assessment (DESSA),* Villanova, Pa.: Devereux Center for Resilient Children, 2009.

Lockwood, J. R., Harold Doran, and Daniel F. McCaffrey, "Using R for Estimating Longitudinal Student Achievement Models," *R News*, Vol. 3, No. 3, 2003, pp. 17–23.

McCaffrey, Daniel F., Bing Han, and J.R. Lockwood, "Using Auxiliary Teacher Data to Improve Value-Added: An Application of Small Area Estimation to Middle School Mathematics Teachers," in Robert W. Lissitz and Hong Jiao, eds., *Value Added Modeling and Growth Modeling with Particular Application to Teacher and School Effectiveness*, Charlotte, N.C.: Information Age Publishing, 2013.

Schochet, Peter Z., *Technical Methods Report: Guidelines for Multiple Testing in Impact Evaluations*, Washington, D.C.: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education, NCEE 2008-4018, 2008.

U.S. Department of Education. (2014). What Works Clearinghouse: Procedures and Standards Handbook (Version 3.0): Institute of Education Sciences.

White, Halbert, "A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity," *Econometrica*, Vol. 48, No. 4, 1980, pp. 817–838.