New Mexico StartSmart K-3 Plus Validation Study

Evaluator's Report

Damon Cann, Mustafa Karakaplan, Margaret Lubke, and Cyndi Rowland
Utah State University

12-22-2015

General Project Background

The StartSmart K-3 Plus project began as an effort to evaluate the effectiveness of the innovative K-3 Plus model as implemented in the State of New Mexico. The K-3 Plus program lengthens the school year by providing an additional 25 days of education during the summer each year prior to grades K-3. Providing summer instruction could be helpful at many points in time. However, these grades may be uniquely important as research suggests that summer learning loss is greatest in the early grade levels, and accumulated summer learning loss over several years can amount to a substantial amount of under-achievement (Alexander, Entwisle, and Olsen, 2007). These losses are most pronounced for students from lower SES groups (Entwisle, Alexander, and Olsen 2001). The logic behind the K-3 Plus program is that it not only minimizes summer learning loss (by putting students in an enriching environment during the summer) but also helps students to experience gains during the summer to propel them forward in their achievement.

Given the substantial achievement gap between low and high SES students in New Mexico, the State of New Mexico began a program known as the K-3 Plus Extended School Year Program. The program is not remedial, but is intended to offer a longer school year to improve achievement and minimize summer learning loss. The program offers 25 additional school days with classes no larger than those in the regular school year. Meals and transportation are provided consistent with the way those services are provided during the school year. Instruction is to be centered on literacy and numeracy and be delivered by certified teachers who have completed professional development in literacy. Teachers are also required to incorporate a parental involvement component, though the exact nature of that component is not specified in statute. At the time we began our study, schools had to have 85% or more of their student body qualify for free or reduced-price lunches (FRL); the threshold has since been reduced to 80%.[1]

A variety of studies support the general effectiveness of summer programs (Borman and Dowling 2006; Downey, Hippel, and Broh, 2004), but it is important to validate these findings by evaluating the specific program as implemented in New Mexico. We have performed a carefully crafted randomized controlled trial of the K-3 Plus program and report on our results here.

Randomized Controlled Trial Structure

The state of New Mexico continued to run its program while we created a parallel program, the StartSmart K-3 Plus program, which mirrors the standards of the state program with one important exception: Students in StartSmart K-3 Plus schools were randomly assigned to either receive the 25 days of summer services plus regular school year services (the intervention group) or to receive regular

[1] In the StartSmart K-3 Plus RCT program, we allowed schools with as low as 70% of students qualifying for FRL to participate in the program. This increases the generalizability of our results while staying true to the intent of the program to target high-need schools.

school year services only (the control group). Thus, differences between the two groups can be attributed to receiving (or not receiving) the program.[2]

Students were recruited in two cohorts, one that began Kindergarten in the Fall of 2011 and a second cohort that began Kindergarten in the Fall of 2012. Both sets of students were followed over four years until the beginning of what would be their 3rd grade year if they had made normal progress in school. If summer class sizes fell below minimum thresholds (8 students), refresh students were added to the classes, but the refresh students are not used in our analyses here (they were added to keep class sizes consistent with those observed in the state-funded program). Nine school districts of varying sizes from across the state of New Mexico agreed to enroll students in the StartSmart study, though two of those districts dropped out of the study after just a single year of participation.[3]

We evaluate students academic achievement in a variety of different outcome domains. We measure expressive vocabulary using the Picture Vocabulary test from the Woodcock-Johnson Tests of Achievement III.[4] We measure reading skills using the Broad Reading cluster score from the Woodcock-Johnson at more advanced levels but use the simpler Letter-Word Identification as our measure at the pre-K and Kindergarten time points. Similarly, we measure math skills using the Broad Math cluster score from the Woodcock-Johnson, but substitute the simpler Applied Problems subtest as our measure at the pre-K and Kindergarten time points. Writing skills are measured using the Basic Writing cluster score from the Woodcock Johnson. We also measured student social skills using the Social Skills Improvement System (SSIS, parent form) and receptive language skills using the Peabody Picture Vocabulary Test (PPVT). Because the PPVT does not have a current Spanish-language equivalent, our sample in that outcome domain is limited to students for whom English was the best language at the time of randomization into the study.[5] All of these measures perform well in terms of reliability and validity. Analyses are completed with the standardized scores produced by the publishers guidelines for each respective assessment tool.

The assessments were administered by trained assessors who had completed a rigorous preparation process. Students were assessed twice each year,

[2] In this report we cover the key aspects of the research design. Additional details are available in our pre-registered research design that accompanies this report on our website.

[3] In these two districts, we were unable to continue to assess either the treatment or the control students. This loss is undoubtedly disappointing, but because it affects both intervention and control group students equally, there is no threat to the integrity of the experimental design.

[4] For this and our other measures based on the Woodcock-Johnson III, we administered the Bateria Woodcock-Munoz for Spanish-speaking students.

[5] Our Spanish-speaking students completed the Test de Vocabulario en Imagenes Peabody (TVIP) as a measure of receptive language, but the scores do not clearly fall on the same metric as the PPVT. The sample of Spanish-speakers at randomization is not large enough to have reasonable statistical power to merit a separate evaluation of Spanish-speaking students alone on this outcome domain.

once in the spring within the first 6 weeks of school and again during the last 6 weeks of the school year. Within each school, all students were tested within 8 days of each other to ensure that students within a school were assessed all at a similar time point within their learning experience. At the same points in time, we asked parents to complete some surveys with some basic information about their child and their family.

Statistical Approach

To test for differences between the intervention and control groups, we use a hierarchical linear modeling framework for multisite randomized trials (see Raudenbush and Liu 2000). In a sense, this approach treats each site as an individual RCT with a planned meta-analysis of the results contributed by each site. The hierarchical model is well-suited for analyzing data like ours where one has units of observation nested within higher-level units; the specific application of such a model for a multi-site RCT is explicated in Raudenbush and Liu (2000). Using the language of Raudenbush and Bryk (2002), we formulate a level 1 model for individual student outcomes on assessments. Within each site *j*,[6] we denote the assessment outcome for student *i* as Y_{ijt},(note that *t* represents beginning of kindergarten or beginning of first grade depending on the analysis) which is a function of a site intercept, β_{0j}, and the site-specific treatment effect, β_{1j}, with X_{ij} as a dichotomous indicator for treatment group membership. In the level 1 model we also control for pre-test performance at baseline (Y_{ijt-1}), gender (where X_{ij2}=1 for females and 0 for males), and maternal education (X_{ij3}=1 for mothers with a high school diploma and 0 otherwise; X_{ij4}=1 for mothers with a college degree and 0 otherwise) as in

$$Y_{ijt} = \beta_{0j} + \beta_{1j}X_{ij1} + \beta_2 Y_{ijt-1} + \beta_{1j}X_{ij2} + \beta_{1j}X_{ij3} + \beta_{1j}X_{ij4} + r_{ij} \qquad (1)$$

where $r_{ij} \sim$ i.i.d. $N(0, \sigma^2)$.

The need for a level 2 model arises because the site intercept and site-specific treatment effect will vary across sites. We specify these site-specific parameters as

$$\begin{aligned} \beta_{0j} &= \gamma_{00} + u_{0j} \qquad (2) \\ \beta_{1j} &= \gamma_{10} + u_{1j} \end{aligned}$$

where γ_{00} is the overall intercept (the grand mean for the control group), γ_{10} is the average treatment effect, and u_{0j} and u_{1j} are bivariate normal site-specific random effects, as in

$$\begin{pmatrix} u_{0j} \\ u_{1j} \end{pmatrix} \sim i.i.d. N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{bmatrix} \tau_{00} & \tau_{01} \\ \tau_{10} & \tau_{11} \end{bmatrix}\right) \qquad (3)$$

[6] A site is defined by the cohort of the student and the school of attendance.

Different outcome domains of achievement may be represented in the dependent variable to evaluate the effectiveness of K-3 Plus on different outcomes of interest. These effects constitute intent-to-treat (ITT) effects as some number of students assigned to the intervention group did not conform to their treatment assignment (e.g. they chose not attend the summer session even though they were intervention students). A handful of control group students also found ways to attend a summer class, though the number of control group crossover students is very small. ITT effects are generally conservative estimates of the effect of treatment on those who actually received the treatment.

Exploratory Results: Kindergarten Readiness

We begin our presentation of estimation results with our data on K-3 Plus and Kindergarten readiness. Recall that our students were assessed in the spring prior to Kindergarten and then received summer services several weeks later. The students were then assessed at the beginning of Kindergarten. For these analyses, the beginning of Kindergarten achievement reflects a students' level of preparedness to begin school. In addition to coefficients and standard errors for these models, the reader will find an effect size (*d*), calculated as the coefficient associated with treatment divided by the standard deviation of the control group. We also present a basic test of baseline equivalence (the difference between the average score for the treatment group at baseline for the analytic sample and the average score for the control group at baseline for the analytic sample divided by the pooled standard deviation at baseline). We follow the What Works Clearinghouse standards that a difference of less than .25 standard deviations between intervention and control groups is acceptable for declaring groups to be equivalent at the baseline time point (assuming one controls for achievement levels at baseline; consider Ho, Imai, King, and Stuart, 2007). Before proceeding to interpretation, we note that in all instances, we observe differences between groups on the baseline assessments that are far less than .25 standard deviations, meeting general standards for baseline equivalence.

We find statistically significant effects of the program on Kindergarten readiness in four out of the six outcome areas: Expressive vocabulary, Reading (measured by letter-word identification), Math (measured by Applied Problems), and Writing. The effect sizes are particularly robust for students on reading outcomes (nearly a third of a standard deviation on the letter-word identification test) and writing outcomes (just over a quarter of a standard deviation). For expressive vocabulary and math (applied problems) the effects are somewhat smaller, with *d* being equal to .092 and .144 respectively. The program does not appear to have statistically significant effects on social skills or receptive language.

Table 1: Exploratory Results for Kindergarten Effectiveness

Dependent Variable:	Expressive Vocabulary	Letter-Word ID	Applied Problems	Basic Writing	Social Skills	Receptive Language
	Coef. (Std. Err.)	Coef. (Std. Err.)	Coef. (Std. Err.)	Coef. (Std. Err.)	Coef. (Std. Err.)	Coef. (Std. Err.)
Treatment	1.562*	4.065*	1.806*	4.610*	.151	.809
Group	(.582)	(.560)	(.437)	(.641)	(.651)	(.489)
Pre-test	.749*	.660*	.592*	.599*	.588*	.716*
	(.021)	(.018)	(.017)	(.018)	(.021)	(.016)
Female	-1.752*	.693	1.034*	1.497*	2.030*	-.293
	(.580)	(.477)	(.428)	(.644)	(.652)	(.490)
Maternal	2.500*	2.734*	1.437*	3.271*	-1.473	.458
Ed.: H.S.	(.740)	(.604)	(.539)	(.806)	(.809)	(.726)
Maternal	5.539*	3.625*	3.701*	5.539*	-.367	1.303
Ed.: College	(1.137)	(.937)	(.850)	(1.266)	(1.238)	(.997)
Constant	21.882*	27.814 *	35.381*	30.993*	40.445*	28.251*
	(1.968)	(1.683)	(1.599)	(1.569)	(2.307)	(1.630)
$\sigma_{Treatment}$	2.97×10^{-8}	2.799	.795	3.32×10^{-6}	3.32×10^{-6}	.0001
	(6.32×10^{-6})	(.568)	(.960)	(.0012)	(.0009)	(.0002)
$\sigma_{Constant}$	1.167	.436	1.324	7.75×10^{-6}	1.035	1.463
	(.566)	(1.438)	(.393)	(1.24×10^{-5})	(.913)	(.417)
Wald X^2	1456.01	1534.44	1457.57	1325.04	799.13	1955.37
	($p < .001$)	($p < .001$)	($p < .001$)	($p < .001$)	($p < .001$)	($p < .001$)
n	1513	1535	1491	1461	1531	1183
d	.092	.312	.144	.271	.009	.055
Baseline Equiv. (Treat - Control)	.074	.073	.128	.095	.011	.024

Notes: Cell entries are coefficients from mixed effects regression with standard errors in parentheses. For coefficients, * denotes $p < .05$, two-tailed. Standard deviations of the random effects on treatment and constant are listed with their standard errors. Effect size (Cohen's *d*) is reported as *d* with the numerator as effect size and the numerator as the standard deviation of the control group at the beginning of Kindergarten.

Confirmatory/Impact Analysis Results: Beginning of 3rd Grade

Our confirmatory contrasts focus on the final time point at which we were able to observe our students. For students who made normal progress in school, this was the beginning of the 3rd grade year (Fall of 2014 for our first cohort and Fall of 2015 for our second cohort).[7] At this point, students in the intervention group were able to have access to the intervention for four summers in addition to regular school year services. Students in the control group had only access to regular school year services. We should note that compliance with group assignment was very good in the control group, with 94% of control group students never receiving summer services. About 5% of control group students found a way to attend one summer and less than 1% attended two summers. No control group student received 3 or 4 summers of the intervention. Compliance was weaker in the intervention group in a sense, with only about 29% of students attending all four years, and an additional 18% attending 3 out of the 4 summers. The attendance tended to be weaker among students in higher grade levels. Because our estimates are for intent to treat effects, the relatively low compliance in the intervention group makes our estimates of program effectiveness quite conservative.

After students receive four years of the intervention and are assessed at the beginning of their 4th year of K-12 schooling, we see that students show some gains in reading and math (about a tenth of a standard deviation) as well as in writing (about .15 standard deviations). However, the program does not show statistically significant effects for Expressive Vocabulary, Receptive Language, or Social Skills at the final time point.

Discussion

Overall, the New Mexico K-3 Plus program displays a measure of promise as an intervention that can improve students performance. Our exploratory results for the Kindergarten year suggest that even just 25 days of summer programming can move students forward in their academic achievement and preparedness for Kindergarten. After four years of the program, the effects of the program are smaller. This could be for a variety of reasons, but our prime suspicion is that the decline in effectiveness is an artifact of students being less likely to actually attend a summer session in the later grades in which the program is offered. We commend an exploration of treatment on the treated effects as a direction for future research to determine whether the low rate of compliance with treatment assignment in the intervention group may account at least partially for the small effect sizes.

Additionally, we have looked here only at the beginning of the year to see what happens with students shortly after they have had the summer program. It is natural to

[7] If a student was retained in grade, we continued to assess them and intervention students were offered summer services. Students who were retained in grade continue to be pooled with students in their cohort regardless of whether students were in the intervention or control group. In essence, treatment means they were eligible to receive 4 summers of services in addition to regular school year services (regardless of whether the student was retained in grade) while control group status means the student was only offered regular school year services (regardless of retention).

wonder whether the programs effects will linger through the course of the school year or if the students are getting pushed forward in the summer only to regress back during the school year.

Whatever the case, the K-3 Plus program is a novel and innovative approach to improving student achievement and is worthy of further study to better understand the nature of the program.

Table 2: Confirmatory Results for Beginning of 3rd Grade

Dependent	Expressive	Broad	Broad	Basic	Social	Receptive
Variable:	Vocabulary	Reading	Math	Writing	Skills	Language
	Coef.	Coef.	Coef.	Coef.	Coef.	Coef.
	(Std. Err.)	(Std. Err.)	(Std. Err.)	(Std. Err.)	(Std. Err.)	(Std. Err.)
Treatment	.400	1.684*	1.374*	2.037*	1.048	.036
Group	(.603)	(.741)	(.671)	(.690)	(.856)	(.555)
Pre-test	.508*	.404*	.669*	.298*	.499*	.609*
	(.022)	(.029)	(.027)	(.020)	(.027)	(.019)
Female	-2.212*	.908	-2.125*	1.956*	2.278*	-.831
	(.586)	(.743)	(.671)	(.691)	(.834)	(.556)
Maternal	1.741*	2.145*	1.857*	1.718	-1.114	.290
Ed.: H.S.	(.773)	(.968)	(.868)	(.901)	(1.079)	(.811)
Maternal	4.017*	2.594*	2.758*	2.150*	-2.236	.082
Ed.: College	(.873)	(1.098)	(.984)	(1.023)	(1.198)	(.885)
Constant	44.081*	55.284*	26.368*	62.721*	53.018*	40.743*
	(2.030)	(2.710)	(2.551)	(1.723)	(2.943)	(1.823)
$\sigma_{\text{Treatment}}$	1.158	3.23×10^{-6}	3.23×10^{-10}	9.60×10^{-10}	1.525	2.06×10^{-7}
	(1.371)	(5.89×10^{-6})	(1.22×10^{-9})	(2.13×10^{-9})	(1.393)	(3.95×10^{-7})
σ_{Constant}	2.282	4.313	2.622	2.485	1.89×10^{-6}	2.305
	(.584)	(.558)	(.551)	(.523)	(3.91×10^{-6})	(.448)
Wald X^2	662.72	229.08	706.80	286.72	349.18	1075.29
	($p < .001$)	($p < .001$)	($p < .001$)	($p < .001$)	($p < .001$)	($p < .001$)
n	1293	1313	1276	1276	1089	1021
d	.030	.111	.091	.148	.065	.003
Baseline Equiv. (Treat - Control)	.076	.065	.129	.109	.018	.011

Notes: Cell entries are coefficients from mixed effects regression with standard errors in parentheses. For coefficients, * denotes $p < .05$, two-tailed. Standard deviations of the random effects on treatment and constant are listed with their standard errors. Effect size (Cohen's *d*) is reported as *d* with the numerator as effect size and the numerator as the standard deviation of the control group at the beginning of Kindergarten.

References

Alexander, Karl L., Doris R. Entwisle, and Linda S. Olson. 2007. "Lasting Consequences of the Summer Learning Gap." *American Sociological Review* 72: 167-80.

Borman, Geoffrey D., and N. Maritza Dowling. 2006. "Longitudinal Achievement Effects of Multiyear Summer School: Evidence from the Teach Baltimore Randomized Field Trial." *Educational Evaluation and Policy Analysis* 28: 25-48.

Downey, D. B., von Hippel, P. T., and Broh, B. A. 2004. "Are Schools the Great Equalizer? Cognitive Inequality During the Summer Months and the School Year. *American Sociological Review, 69,* 613–635.

Entwisle, Doris R., Karl L. Alexander, and Linda S. Olson. 2001. "Keep the Faucet Flowing." *American Educator* 25(3): 10-15.

Ho, D., Imai, K., King, G., and Stuart, E. A. 2007. Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. *Political Analysis*, 15(3), 199– 236.

Raudenbush, Stephen W. and Anthony S. Bryk. 2002. *Hierarchical Linear Models: Applications and Data Analysis Methods.* Sage: Thousand Oaks, CA.

Raudenbush, Stephen W. and Xiaofeng Liu. 2000. "Statistical Power and Optimal Design for Multisite Randomized Trials." *Psychological Methods*, 5(2), 199-213.