Using Entropy Balancing to Reduce the Effects of Selection Bias in Afterschool Studies: An Example in Studying the Relationship between Intensity of Afterschool Program Participation and Academic Achievement

Denise Huang, Seth Leon & Deborah La Torre

Abstract: Every since the enactment of the No Child Left Behind Act (2001) in the United States, achievement gains resulting from afterschool participation have been of particular interest. However, findings have been inconsistent. The challenge for researchers is partly due to the wide variation of program goals, difficulty in obtaining valid control groups, difficulty in obtaining clean records of data, the high transience rates of the students, and in particular, the failure to differentiate among the dosage students receive and the inherent potential of selection bias in the afterschool population. This study draw on a large dataset and allows for the analysis of effects over the course of several years. Using LA's BEST afterschool program as an example, this study employed advanced methodology to reduce selection bias in examining the relations between afterschool program participation and academic achievement.

Keywords: afterschool program, participation, academic achievement, selection bias, LA's BEST

Introduction

Since the turn of the century, interest and funding in afterschool programs has increased significantly. For example, California increased its yearly budget for afterschool programs from 120 to 550 million during the 2006–07 fiscal year (California AfterSchool Network, 2007). As a result, funders and policymakers are demanding greater accountability of programs. In particular, with the enactment of the No Child Left Behind Act in 2001, achievement gains resulting from afterschool participation have been of particular interest (Lauer, Akiba, Wilkerson, Apthorp, Show, & Martin-Glenn, 2006). However, while many researchers consider afterschool programs to be a potentially powerful resource to achieve this goal, the reported findings on academic outcomes tend to be mixed (TASC, 2005; Vanderhaar & Muñoz, 2006). This is thought to be due to a wide variety of reasons including the wide variation of afterschool program goals, difficulty in obtaining valid control groups, access to clean

 https://doi.org/10.3224/ijree.v5i1.01

program records, and high transience rates among staff and students (Lauer et al., 2006). In particular, we believe that studies of these programs are impacted by the inherent potential of selection bias in the afterschool population and the failure of many researchers to differentiate among the dosage (participation hours or days) that student participants receive (Lauer et al., 2006).

With these issues in mind, the present study furthers the afterschool research through the use of an entropy balancing technique to reduce self-selection bias among a large student sample. This study employed this advanced method to examine the longitudinal effect of dosage on students' academic outcomes over a period of four years. Accordingly, the main research question for this study is as follows: Do the achievement outcomes of LA's BEST students' vary as a function of their different intensity levels of afterschool participation?

Review of the Literature

Dosage is a critical factor to examine when assessing the effect of an intervention. This is because an examination of dosage enables researchers to determine whether participants are receiving a sufficient treatment in order to demonstrate an effect. Even though dosage, or intensity of participation, is important to determining program success, it has only recently been examined in the literature on afterschool programs. In general, these studies found a positive relationship between intensity of participation and positive student outcomes. For instance, Frankel and Daley (2007) found that higher afterschool attendance was associated with higher academic achievement. In addition, Goldschmidt, Huang and Chinen (2007), found that medium (10–14 days per month) and high attendance (15 or more days per month) in an afterschool program was associated with lower juvenile crime rate. Multiple studies also found a relationship between afterschool attendance intensity and higher day school attendance (Frankel & Daley, 2007; Jenner and Jenner (2007); Huang, Gribbons, Kim, Lee, & Baker, 2000; Welsh, Russell, Williams, Reisner & White, 2002; Munoz, 2002).

Thus, in reviewing research on participation and outcomes in afterschool programs, it appears that many studies that claim positive outcomes reported academic improvement in students with a higher dosage of afterschool participation, while those that reported null or negative findings more often looked at participants of afterschool programs as an aggregated group. As such, we believe it is important that those who study afterschool program effects to consider examining "dosage" or intensity level.

Reducing Selection Bias

Another frequent critique of afterschool studies is selection bias (Hollister, 2003; Little & Harris, 2003; Scott-Little, Hamann, & Jurs, 2002). Because afterschool program participation is voluntary, students (or their parents) self-select themselves

into participation and non-participation groups. In comparing participating students to non- participating students in the same school, there are inherent biases that researchers need to balance or control in order for the findings to be valid.

Furthermore, while the U.S. Department of Education (2003) has emphasized the importance of using experimental designs with control groups in educational research, reaching this "gold standard" is difficult in afterschool programs due to social contexts. Moreover, it is often difficult and potentially unethical for most afterschool programs to randomize their participants unless they are grossly oversubscribed. More specifically, unless programs have many more applicants than available spaces, random assignment would mean refusing to accept some students into the program so that they could serve as controls. Students who are refused enrollment may end up unsupervised and without the homework help they desperately need.

As a result, many studies lack a true experimental design or control group. Thus, most studies in this field are quasi-experimental, with researchers using a comparison group and making use of statistical controls. In these quasi-experimental studies, one needs to be cautious when inferring causality. With this in mind, the present study reduces self-selection bias by removing pre-existing category differences using entropy balancing. This method was employed because it has been shown in simulations and empirical applications to lower approximation error and reduce model dependency (Hainmueller, 2011). More specifically, use of this method enabled us to remove differences in observed student background characteristics between those who attended the afterschool program and those who did not. It also enabled us to attribute differences in achievement outcomes to treatment dosage with more confidence.

The sample base for this study consisted of participants in an afterschool program called Los Angeles' Better Educated Students for Tomorrow (LA's BEST). This program was selected because it serves students in a large school district and it shares many of the common features of quality afterschool programs that serve urban, low-income, and low-performing schools. Additionally, the student demographic profiles for LA's BEST are very similar to the national profiles of urban afterschool participants. Thus, inferences from this study can be generated to other urban afterschool programs serving similar populations. First, a brief description of the LA's BEST program is provided.

The LA's BEST Program

LA's BEST was first implemented in the fall of 1988. The program operates under the auspices of the Mayor of Los Angeles, the Superintendent of the Los Angeles Unified School District (LAUSD), a board of directors, and an advisory board consisting of leaders from business, labor, government, education, and the community.

LA's BEST seeks to provide a safe haven for at-risk students in neighborhoods where gang violence, drugs, and other types of anti-social behaviors are common. The program is housed at selected LAUSD elementary schools and is designed for students in kindergarten through fifth or sixth grade, depending upon the school. The LA's BEST sites are chosen based on certain criteria, such as low academic performance and their location in low-income, high-crime neighborhoods. For opti-

mal program success and to ensure buy-in from the principals and the school staff, the school principals have to write an official letter of request for the program to be placed in their school site.

LA's BEST is a free program open to all students in the selected sites on a first-come, first-served basis. Students who sign up for the program are expected to attend five days a week in order to reap the full benefits of the program. Program offerings include academic assistance, enrichment, and physical activities. At the time of this study, LA's BEST served a student population of approximately 34,000 with about 80% Hispanic and about 12% African American elementary students. English Learners comprised at least half of the student population at most sites. Of this population, the majority's primary language was Spanish, while the other percentage of the English Learner population was composed of those whose first language was of Asian/Pacific origin.

Cognitive beat and homework beat	Recreation beat
Intellectual development: • **Future aspirations** – through high expectations, activities that build self-reliance, value of education, collaboration, and critical thinking. • **Love of learning** – through active participation, exploration, and engaging research-based activities. • **Responsibility and positive work habits** – through emphasis on the importance of completing assignments, teaching learning strategies and study skills, and providing a learning climate that enforces positive attitudes towards school. • **Self-efficacy** – through guided experiences, challenging activities, and relationship building between staff and students.	Physical and social-emotional development: • **Healthy lifestyle** – through curricula and activities that promote drug and gang prevention, healthy eating habits, and plenty of exercise. • **Respect for diversity** – through role modeling and curricula that enhances awareness and responsibility to each other within their diverse community. • **Sense of community** – through providing students with opportunities to participate in community-sponsored events, volunteer in community assignments, and offering field trips to local businesses and organizations. • **Sense of safety & security** – through providing students with a safe and nurturing environment. • **Social competence** – through demonstrating and enhancing students' respect for self and others, and providing students with opportunities to form friendships and develop trust and respect with peers and adults.

Figure 1. LA's BEST 3.5 Beat Structure

Since its inception in 1988, LA's BEST has adapted and updated their goals in response to educational policies, research, and theory. Over the years, the program has moved past its initial emphasis on providing a safe environment and educational enrichment to an emphasis on the development of the whole-child. In developmental theory, a whole-child curriculum is one that cultivates the development of students' intellectual, social, and emotional well-being so that children can achieve their full potential (Schaps, 2006; Hodgkinson, 2006). As shown in Figure 1, at LA's BEST,

their 3.5 beats focus on the whole-child by emphasizing students' intellectual, social-emotional, and physical development.

To summarize, the mission of LA's BEST is to provide elementary age students with a physically and emotionally safe setting during the afterschool hours that is engaging and connects to the school and broader community. And, most importantly, provides students with access to extra-curricular activities, challenging academic enrichment, and qualified, caring adults (see LA's BEST, n.d.).

Study Design and Methods

This study employs a quasi-experimental design that consists of a longitudinal sample of both academic and LA's BEST program attendance data. The sample was comprised of two cohorts of students who had no LA's BEST participation during second grade (2005-06 and 2006-07). The students in each cohort were then followed from third through fifth grade, using an entropy balancing method. This method enabled us to model the sequential treatment status of the students that varied across time. Each model presented defines treatment status based on student dosage (intensity of attendance) in the program in a given year. This was done in order to remove any existing differences in the observed student background characteristics across treatment status. Finally, hierarchical growth modeling was applied to academic outcomes with specific effects of interest estimated. A non-response weight was also included to adjust for missing data.

Two benefits were gained by utilizing the longitudinal nature of the data to follow students' academic development over time. First, it allowed the study to move beyond traditional pre/post analysis, which is limited by data requirements and explanatory possibilities (Rogosa, Brandt, & Zimowski, 1982; Raudenbush & Bryk, 2002). The study was able to employ growth-modeling techniques to examine individual trajectories (Rogosa et al., 1982) and had more flexible data requirements. Second, we were able to adapt an approach developed by Hong & Raudenbush (2008) to study the effects of time-varying treatments on student achievement.

Defining the Study Sample

The basis for this study sample is the LAUSD student database that the research team has collected and stored since the 1992–93 school year. The first step in constructing a study sample is to generate a sampling frame. This task was accomplished by going back through the historical records and tracking four years of background and California Standards Tests (CSTs) achievement data for the students in the two cohorts. Students who were in second grade during the 2005-06 and 2006-07 school years who did not participate in LA's BEST at baseline and who had complete data throughout the study period (e.g., afterschool attendance, achievement scores, day school attendance, behavior ratings, etc.) were included in these cohorts. Since a recent study reported that self-discipline in students is a predictor of academic abilities

(Ponitz, McClelland, Matthews, & Morrison, 2009), the behavior ratings used were limited to five student self-discipline items (i.e., follows direction, accepts and respects authority, shows dependability, take responsibility, and exercises self-control). Furthermore, because of the expansion of the LA's BEST program across its approximately 20-year history, the 2005–06 cohort included students from 148 schools, while the 2006–07 cohort included students from 168 schools.

Examination of student attendance patterns indicates that students participate in afterschool programs with varying regularity. Therefore, it is necessary to set criterion to measure the students' dosage. To accomplish this, attendance levels were set for the treatment students based on the average number of days students participated per year. These included the following: (1) any attendance, (2) 2 days per week or a minimum of 72 days per year, (3) 2.5 days per week or a minimum of 90 days per year, (4) 3 days per week or a minimum of 108 days per week, (5) 3.5 days per week or a minimum of 126 days per week, and (6) 4 days per week or a minimum of 144 days per year.

Using the English language arts sample as an example, Figure 2 illustrates the manner in which students were included in the various models. Approximately 35,000 students had valid data during the baseline years for the two cohorts. Of these students, those who were enrolled in LA's BEST at baseline were excluded. This step was necessary so that all treatment and control students in the study had the same treatment status at baseline. Additional students were excluded from the sample due to missing outcome or background data during third, fourth, or fifth grade. As a result, approximately 12,500 of the students across these two cohorts received some treatment.

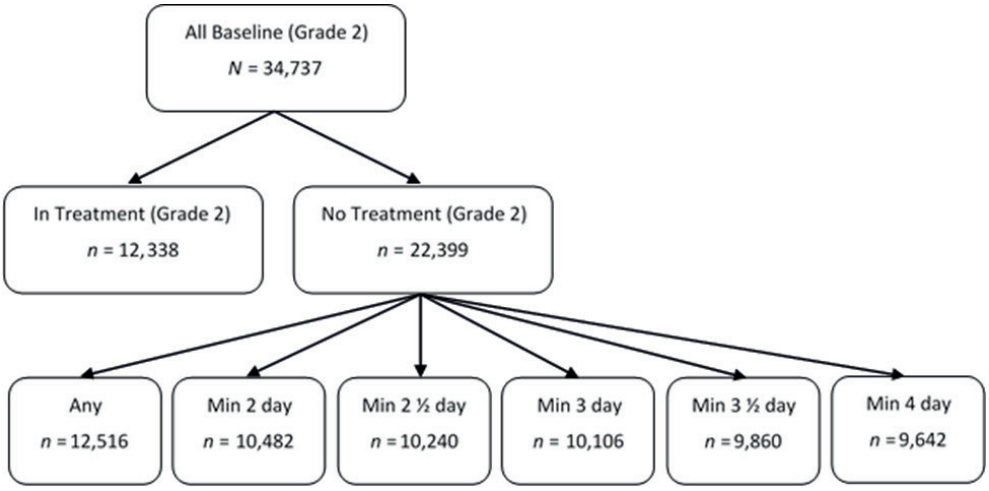

Figure 2. English language arts sample for the two cohorts

Non-Response Data

For the purposes of this study, non-response data is defined as students dropped from the analysis due to either missing data or having attended the LA's BEST program in second grade. The day school attendance and tardiness covariates in this adjustment are omitted because close to 5,000 students were missing data on these indicators at baseline. The day school attendance and tardiness covariates in the primary analyses were included due to their potential connection to attendance in afterschool programs and because they do not add substantially to missing data in the follow-up for third, fourth, and fifth grade.

Means for the baseline covariates used to adjust for non-response are presented in Table 1 for the original sample and for students comprising the math outcome sample. Although no single covariate was strongly associated with non-response, eight of the ten did have some level of significant association. The strongest association with non-response occurred with teacher ratings of student behavior. Students included in the analyses had slightly better baseline behavior ratings than those not included. Non-response was adjusted by creating a predicted probability of non-response using logistic regression entering the ten covariates in Table 1. The predicted probability is later used as a final adjustment weight (see growth model section).

Table 1. Means of Baseline Covariates for Original Sample and Valid Response with Effect (Eta)

	Original *n* = 34,737	Non-Response *n* = 22,221	Any Participation *n* = 12,516	
	Mean	Mean	Mean	Eta
Grade 2 Limited English Proficiency (1 = yes, 2 = no)	0.575	0.564	0.596	0.032**
Grade 2 Cohort (1 = 2005-06, 0 = 2006-07)	0.475	0.492	0.443	0.048**
African American (1 = yes, 2 = no)	0.091	0.105	0.068	0.062**
Hispanic (1 = yes, 2 = no)	0.807	0.794	0.831	0.044**
Parent- Some College (1 = yes, 2 = no)	0.169	0.171	0.166	0.006
Parent- HS Graduate (1 = yes, 2 = no)	0.195	0.192	0.201	0.011*
Parent- Not HS Graduate (1 = yes, 2 = no)	0.289	0.281	0.303	0.023**
Male (1 = yes, 2 = no)	0.507	0.509	0.502	0.008
Grade 2 Behavior Rating (z-score)	0.003	-0.068	0.128	0.094**
Grade 2 Math CST (z-score)	-0.112	-0.152	-0.041	0.056**

Controlling for Existing Population Differences

Students who attend LA's BEST self-select into the program rather than being randomly assigned to attend. Thus, there are likely to be differences in observed data between those who attend the program and those students in the same schools who do not attend. In observational studies, matching and propensity methods are often used by researchers to improve the balance in observed covariates between treatment and control subjects (Hainmueller, 2011; Ho et al., 2007; Sekhon, 2009). This preprocessing step is often approached using logistic regression to estimate the probability that a subject would be in the treatment group. The propensity outcome is then used to create balance among the student background characteristics. This process can be done using matching, stratum, or weighting techniques.

At present there is no accepted consensus concerning which approach to use for preprocessing observational data. Furthermore, one common concern is that the most commonly used methods do not directly or necessarily create balance among covariates (Hainmueller, 2011). This requires the researcher to check carefully that covariate balance has been achieved with a correctly specified propensity model. This can be a time consuming process with no guarantee that covariate balance will be achieved.

For this study, an entropy balancing method was used to preprocess the data. Entropy balancing has several important advantages over propensity methods. The first is that entropy balancing directly balances covariates to preconditions (moments) set by the researcher. If the model converges, then balance has been achieved. Unlike propensity matching methods, and similar to inverse weighting methods, entropy balancing re-weights each observed case which is useful in longitudinal studies. In addition to directly balancing the covariates, this approach includes a second step in which the weights are refined, with large weights being reduced to minimize the variance in the analyses that follow.

Use of balancing covariates. Covariates used in preprocessing observational data should either be static indicators such as gender or ethnicity, or they should be measured prior in time to the treatment indicator on which balance is desired. This ensures that the covariates are not affected by the treatment. Covariates should be included when it is reasonable that they may simultaneously influence selection into treatment and the outcome measure (Caliendo & Kopeinig, 2005).

In this study, it was necessary to balance treatment and control populations in third, fourth, and fifth grade. A set of baseline covariates in second grade were used at each of these grade levels. These baseline covariates included the following: gender, ethnicity, language proficiency status, parent education, student behavior rating, day school attendance %, day school tardy %, CST score, and an indicator of the cohort in which the student belonged. A subset of the baseline indicators that were time varying were included whenever they were observed prior to treatment grade level, as were all prior treatment indicators. For example, to balance the covariates for fourth grade treatment, all baseline covariates were included, the third grade time-varying covariates (i.e., student behavior rating, day school attendance %, day school tardy %, and CST score), and the third grade treatment indicator. Each covariate was entered at the student-level and as a difference from the school mean.

This step was taken to ensure that differences at the school-level that might influence the likelihood of future student attendance in the program would be balanced across treatment and control populations.

Entropy balancing software (available in R) functions to re-weight the control group while keeping the treatment group un-weighted. This produces a weight for analyses that seek to determine the average treatment effect on the treated (ATT). The pseudo-sample necessary for our sequential treatment growth model methodology requires an average treatment effect (ATE) weight. Using entropy balancing, this study separately created a weight for the control and for the treatment groups that balanced each to the total sample (treatment plus control). Weights were normalized to treatment and control original sample numbers so that the mean weight for each of these groups was equal to one. The result was a weighted sample for ATE in which the covariates were balanced across treatment and control.

Entropy Balancing Results

Prior to weighting there were many covariates with significant differences between treatment and control groups making entropy balancing necessary. The prior treatment indicators had the largest association with treatment selection and other individual covariates generally had a small association with treatment selection.

Significant associations between the student-level covariates and treatment selection in third, fourth, and fifth grade can be found in Table 2. Teacher ratings of prior student behavior were consistently lower among treatment students than among control students. This was likely related to student targeting practices at the different afterschool sites. Prior math and ELA CST outcomes were not associated with treatment status in the samples that did not include students who attended LA's BEST an average of less than two days per week. In the models that did include these lower attending students, prior math and ELA CST scores were slightly higher for control students than for treatment students. As might be expected, prior day school attendance differences between treatment and control student were present in the models that require some threshold of attendance intensity. Similar to the math outcomes, prior day school tardy % was more likely to be associated with treatment status in the samples that did include the lower attending students. In these samples, treated students were more likely to be tardy in past school years, which could also be the result of afterschool program targeting. Males were increasingly more likely to be represented in the control than treatment population in fourth and fifth grade. Regarding ethnicity, more African Americans were in the treatment than in the control in third grade. Conversely, fewer Hispanics tended to be in the treatment than in the control.

Table 2. Summary of Significant Occurrences and Largest Effect Sizes (Eta) Across Six Attendance Models

Covariates	Grade 3	Grade 4	Grade 5
Grade 2 Limited English Proficiency (1 = yes, 2 = no)	(5,.036)	(4,.030)	(1,.019)
Grade 2 Cohort (1 = 2005-06, 0 = 2006-07)	(2,.029)	( )	( )
African American (1 = yes, 2 = no)	(5,.030)	(1,.019)	( )
Hispanic (1 = yes, 2 = no)	(5,.040)	(1,.020)	( )
Parent- Some College (1 = yes, 2 = no)	( )	( )	( )
Parent- HS Graduate (1 = yes, 2 = no)	( )	( )	( )
Parent- Not HS Graduate (1 = yes, 2 = no)	(2,.024)	(1,.023)	(5,.027)
Male (1 = yes, 2 = no)	(1,.018)	(5,.031)	(6,.051)
Grade 2 Behavior Rating (z score)	(6,.056)	(6,.054)	(6,.053)
Grade 2 Day School Attendance (%)	(3,.027)	(5,.043)	(5,.044)
Grade 2 Day School Tardy (%)	(2..035)	(1..028)	(3..026)
Grade 2 Math CST (z score)	(2,.046)	(2,.043)	(2,.040)
Grade 2 ELA CST (z score)	(1,.034)	(1,.042)	(2,.031)
Grade 3 Behavior Rating (z score)		(6,.059)	(6,.059)
Grade 3 Day School Attendance (%)		(5,.059)	(5,.040)
Grade 3 Day School Tardy (%)		(1,.032)	(2,.030)
Grade 3 Math CST (z score)		(2,.033)	(2,.043)
Grade 3 ELA CST (z score)		(1,.039)	(2,.046)
Grade 3 Treatment (1 = yes, 2 = no)		(6,.630)	(6,.470)
Grade 4 Behavior Rating (z score)			(6,.022)
Grade 4 Day School Attendance (%)			(5,.052)
Grade 4 Day School Tardy (%)			(2,.021)
Grade 4 Math CST (z score)			(2,.043)
Grade 4 ELACST (z score)			(2,.046)
Grade 4 Treatment (1 = yes, 2 = no)			(6,.767)
Grade 3 and 4 Treatment (1 = both years, 2 = no)			(6,.543)

Entropy balancing was employed to attain weights that balanced covariates across treatment and control separately for third, fourth, and fifth grade. This was done for the ELA and math samples at each of the six program intensity thresholds. In each case, the balancing method converged within tolerance. As a result, after weighting the mean difference for each covariate between treatment and control was essentially equal to zero, and the p-values of t-tests comparing the means equaled one.

It is possible for entropy balancing to converge within tolerance, but still result in some cases with very large weights. When this occurs, it indicates thin support

in the control population for certain covariate combinations in the treated population. Because of this, the largest weights were examined relative to the respective treatment and control populations to determine if extreme weights were a problem in our pseudo-samples. Prior work has suggested that researchers should consider trimming cases when any weight exceeds four to six percent of the sample (Huber, Lechner & Wunsch, 2010). The largest weights in our pseudo-samples occurred in the treated population when the fifth grade sample program intensity was restricted to a minimum of four days per week (144 days). This occurred because the prior treatment helped strengthen the prediction of fifth grade treatment, and because the program intensity restriction reduced the size of the treatment sample. Despite this, the largest weights only represented 2.2% of the ELA treatment sample and 2.4% of the math treatment sample under the four days per week restriction.

Unlike propensity score methods entropy balancing does not produce a prediction of the likelihood of treatment. This method does, however, enable one to infer how well the covariates predict treatment. Control cases with an increasing weight suggest a higher likelihood of treatment. Conversely, treatment cases with an increasing weight suggest a lower likelihood. This study placed the weights on a natural log scale (control cases = -ln(weight), treated cases = ln(weight)) and examined the area under the Receiver Operating Characteristic (ROC) curves to gain a rough understanding of how well the covariates predicted treatment at each grade level under the various program intensity thresholds. Table 3 presents the results for any treatment and those meeting the highest attendance threshold of a minimum of four days per week. After applying this approach, it was clear that treatment in third grade was only weakly predicted by the available covariates. Treatment prediction became somewhat stronger in fourth and fifth grade as the prior treatment indicators were included as covariates. This suggests that our model results may have been vulnerable to an unmeasured covariate that was associated both with the likelihood of treatment and the outcome variable. This vulnerability was strongest in third grade and in the models where no restriction was placed on program attendance intensity.

Table 3. Area under Receiver Operating Characteristic (ROC) Curve

	ELA		Math	
Level of LA's BEST participation to define treatment	Any Treatment	Minimum 4 days per week	Any Treatment	Minimum 4 days per week
Grade 3	0.523	0.549	0.502	0.540
Grade 4	0.557	0.656	0.579	0.699
Grade 5	0.648	0.795	0.640	0.794

Analysis: HLM Growth Modeling

To examine the effects of the afterschool program on achievement and achievement growth, we employed an HLM design that has the advantage of directly modeling growth trajectories (Raudenbush & Bryk, 2002). This type of analysis allows flexible specification of the covariance structure at every level of the analysis for this study (Snijders & Bosker, 1999). This study took advantage of this flexibility by allowing the treatment effect to vary across schools.

As was previously noted, one of the aims in this study was to examine the potential causal effects of afterschool program attendance dosage on achievement growth. The study's data sample allowed the examination of this question to be undertaken longitudinally with treatments and covariates varying across time. Afterschool program attendance dosage can be conceived of as relating to both the intensity of attendance in a given year as well as the pattern and consistency of intensity across time. Hong & Raudenbush (2008) published a paper that outlined an approach for exploring causal effects with time-varying treatments within an educational setting. This study's approach follows theirs closely while adapting for differences in available data and concepts of treatment. In this study, treatment selection was conceived as occurring primarily at the student-level rather than at the level of the classroom. Even though this study lacked data connecting students to teachers and classrooms, there are data on three years of outcome and treatment opportunities after baseline.

Defining Specific Treatment Effects of Interest

The availability of three years of outcome and treatment opportunities after baseline led to many potential effects. For example, there was one effect of treatment on the third grade outcome, three possible effects on the fourth grade outcome, and six more on the fifth grade outcome. To reduce the complexity of interpretation this study collapsed the potential effects into four categories. The first effect moving into treatment was named MIT. This category includes all students who had treatment in a given year after having no treatment in the previous year. The MIT effect was found in third, fourth, and fifth grade since all students in the samples were selected based on having no treatment in second grade. The second effect moving out of treatment were named MOT. This category requires no treatment in a given year after receiving treatment in the previous year. This effect could only occur in the fourth and fifth grade outcomes. MOT collapses the third grade effect on fourth grade as well as the fourth grade effect on fifth grade. Our third category, two years of consecutive treatment (CYT2), requires students to receive treatment in a given year as well as the previous year. This effect was coded when treatment was present in both third and fourth grade for the fourth grade outcome, and if treatment was present in both fourth and fifth grade for the fifth grade outcome. Finally, a category was created for those students who received three years of consecutive treatment (CYT3). This effect was coded when students received treatment in all three school years and applied to the fifth grade outcome.

Examples of how the treated samples were distributed across the groups that represent the specific effects of interest are displayed in Table 4. Results are categorized by grade level for all of the samples. Furthermore, since the sample sizes differed by only a small number of cases, only the results for the math sample and not for ELA were represented. As can be seen, all third grade students were classified as moving into treatment. In contrast, fourth grade included samples of students moving into treatment, out of treatment, as well as those with two consecutive years. Likewise, fifth grade included students in all four categories. As a result, the sample sizes of the treatment students dropped substantially as the restrictions on program attendance were increased. Nevertheless there was a reasonable distribution across the effects such that no single effect predominated.

Table 4. Students Treatment Status Across Grade Levels – Math Sample

Level of LA's BEST participation to define treatment	Total treated	Moving into treatment	Moving out of treatment	Two years consecutive treatment	Three years consecutive treatment
	Any Participation (n = 12,516)				
Grade 3	1,813	1,813	--	--	--
Grade 4	2,030	980	763	1,050	--
Grade 5	2,065	821	786	533	711
	Minimum 3 day weekly average (108 days; n = 10,063)				
Grade 3	523	523	--	--	--
Grade 4	693	302	132	391	--
Grade 5	783	228	138	241	314
	Minimum 4 day weekly average (144 days; n = 9,655)				
Grade 3	306	306	--	--	--
Grade 4	429	198	75	231	--
Grade 5	494	135	70	169	190

Three-Level HLM Growth Model

The HLM analysis employed was based on a three-level model. At Level 1, the standardized achievement score was modeled to be predicted by time (school year) and the treatment effects. This model includes six coefficients for each student, including an intercept and a slope, and the four previously defined treatment effects of interest. The intercept at this level is the student's status at the first time point.

$$ZCST = \pi_0 + \pi_1(Time) + \pi_2(MIT) + \pi_3(MOT) + \pi_4(CYT2) + \pi_5(CYT3) + e$$

Level 2 was modeled to account for student-level effects. Only the student-specific intercepts and growth rates were allowed vary randomly over Level 2.

$$\pi_0 = \beta_{00} + r_0$$
$$\pi_1 = \beta_{10} + r_1$$
$$\pi_2 = \beta_{20}$$
$$\pi_3 = \beta_{30}$$
$$\pi_4 = \beta_{40}$$
$$\pi_5 = \beta_{50}$$

At Level 3, the school level was included in the model. The intercept, slope and all treatment effects were allowed to vary randomly over this level.

$$\beta_{00} = \gamma_{000} + \mu_{00}$$
$$\beta_{10} = \gamma_{100} + \mu_{10}$$
$$\beta_{20} = \gamma_{200} + \mu_{20}$$
$$\beta_{30} = \gamma_{300} + \mu_{30}$$
$$\beta_{40} = \gamma_{400} + \mu_{40}$$
$$\beta_{50} = \gamma_{500} + \mu_{50}$$

Two separate models were conducted: one for math and one for English language arts. In these models, Level 1 represented time nested within students. There were four time points for each achievement model, with achievement at each time point serving as the outcome.

Applying weights. It has been shown that a weight that is inversely related to the probability of treatment (IPTW) can be applied to approximate data from a random sample (Robins, Herna´n, & Brumback, 2000). Using entropy balancing a weight was created that balanced observed covariates, including all prior treatment combinations, across the treatment and control populations in third, fourth, and fifth grade. The goal of this weight, like with IPTW, was to create a pseudo-sample that approximates data from a random sample. Hong & Raudenbush (2008) have shown that the IPTW method in single level settings can be applied to a multilevel educational setting. Strong sequential ignorability is defined so that treatment assignment at a given time point is independent of all potential outcomes given past observables. The weight that applies to sequential settings is conditional and cumulative.

Within the study's analyses, the entropy balancing weights were defined as follows: The weight for third grade in a sequential setting (*sw3*) is simply equal to *w3*; The weight for fourth grade in a sequential setting (*sw4*) is equal to *w3*w4*; and, the weight for fifth grade (*sw5*) is equal to *w1*w2*w3*. In addition a non-response weight was created, which was inversely proportional to the estimated probability of having valid data given the observed baseline covariates. The final weight is the product of the sequential treatment weight and the non-response weight.

HLM Results for English Language Arts and Math Achievement

The following presents the results from the HLM models for English language arts and math achievement.

English language arts achievement results. Table 5 presents the results from the three-level HLM growth models for English language arts. All significant effects were found in the models in which subjects with less than 126 days of afterschool program dosage were not restricted from the analyses. For each of the three models including subjects with less than 126 days of afterschool program dosage the moving into treatment effect was significant. Each of these significant moving into treatment effects were in the negative direction and had very small effect sizes. There was no clear trend that could be attributed to increased program dosage in a given year or across time for students with consecutive treatments. Because the baseline covariates did not strongly predict treatment status in third grade, some unmeasured confounder could potentially be responsible for the significant negative findings. If program participation was actually leading to reduced performance in English language arts we would expect the negative findings to become stronger in the analyses that focus on students' receiving higher participation. Further work is need to derive any meaningful interpretations from these results.

Table 5. Estimated Impact of LA's BEST Intensity of Participation on ELA CST

	Estimated Treatment effects in SD units			
Level of LA's BEST participation to define treatment	Moving into treatment	Moving out of treatment	Two years consecutive treatment	Three years consecutive treatment
Any Participation	-0.034 **	0.007	-0.056**	-0.043
Minimum 2 day weekly average (72 days)	-0.043 **	-0.032	-0.114 **	-0.079
Minimum 2 1/2 day weekly average (90 days)	-0.042 *	-0.029	-0.098**	-0.119**
Minimum 3 day weekly average (108 days)	-0.035*	-0.040	-0.081	-0.086
Minimum 3 1/2 day weekly average (126 days)	-0.028	-0.053	-0.075	-0.045
Minimum 4 day weekly average (144 days)	-0.034	-0.078	-0.078	-0.011

Math achievement. Table 6 presents the results from the three-level HLM growth models for Math. For the model including any participation, the moving into treatment effect was significant ($p > 0.05$) in the negative direction with a very small effect size. There was a trend associated with increased program dosage for students with two years of consecutive treatment as well as three years of consecutive treatment. Students who attended the program with three years of consecutive treatment began to exhibit a significant effect ($p > 0.05$) when their attendance dosage was a minimum

of 108 days in each grade. The effect was also significant with a larger effect size when the attendance dosage was a minimum of 144 days in each grade.

Table 6. Estimated Impact of LA's BEST Intensity of Participation on Math CST

	Estimated Treatment effects in SD units			
Level of LA's BEST participation to define treatment	Moving into treatment	Moving out of treatment	Two years consecutive treatment	Three years consecutive treatment
Any Participation	-0.031 *	0.006	-0.003	0.000
Minimum 2 day weekly average (72 days)	-0.027	-0.029	-0.013	0.000
Minimum 2 1/2 day weekly average (90 days)	-0.037	-0.067	0.041	0.045
Minimum 3 day weekly average (108 days)	-0.045	-0.058	0.078	0.100*
Minimum 3 1/2 day weekly average (126 days)	-0.041	-0.013	0.114	0.109
Minimum 4 day weekly average (144 days)	-0.024	-0.035	0.152	0.207 **

Summary. Results concerning language arts in terms of dosage and effects are inconclusive and require further study. Negative findings that were present when the treatment definitions included less than regular attendance were not maintained when treatment was defined as requiring consistent and regular attendance. Future studies can examine this inconsistence in more details.

In contrast, results of the analyses for math provided evidence that regular attendance in the afterschool program for a period of three consecutive years may lead to positive achievement growth. This finding was first significant at the 108 days per year (three days per week) threshold, and the size of the effect increased at 144 days per year (four days per week). When program attendance were causally related to math achievement, one could expect to find these results. Given the intrigue methodological steps that have been taken, this study concludes that the most plausible explanation for these results is due to program dosage effect. However, this study cautiously stops short of making a causal inference since there is still some potential for the strong ignorability assumption to be violated by an unmeasured confounder. Similar future studies can add support and strengthen the claim in this study.

Discussion and Conclusion

The literature provides evidence that quality afterschool programs can teach students academic and social skills, help them avoid anti-social behavior, and contribute to academic resiliency (Bradshaw et al., 2013; Durlak et al., 2010; Maynard et al., 2013, McKinsey & Company, 2009). However, sufficient exposure to effective afterschool environments is necessary in order for students to reap the benefits. At the same time, while it seems to be necessary to look at the intensity of participation (dosage) as a contribution to student outcomes, in order to have valid findings it is also important to control for the selection bias that is inherent in the field of afterschool research. This study set out to reduce a research gap by using rigorous methodology to study the effects of dosage on students' academic outcomes. It extends the current literature on the impact of afterschool programs in three key ways: first, the analyses explicitly modeled sequential program attendance and achievement outcomes longitudinally for four years; second, we defined program dosage in two dimensions (within a given year and across years); third, we used a large sample of roughly 35,000 students. Finally, we took steps to apply an entropy balancing technique and establish a valid study pseudo-sample from which we could generate valid inferences.

Currently, there are very few afterschool studies that have involved such a large study sample. This large sample size added substantial strength to the findings in this study. Furthermore, with all the careful, meticulous, and intentional methodologies, examinations, and interpretations, the findings in this study add support to the notion that regular attendance is necessary to reap benefits in math achievement. Therefore, after school programs can enhance their efforts in encouraging students to participate regularly so that they can reap the program benefits. Since this study focuses on math and English Language Arts achievement future studies can also elaborate more on other social outcomes and the dosage effect.

Implication on Methodology

Studies of afterschool programs typically are designed to compare participants and non-participants without careful examination of the dosage effect. Consequently, participants may attend one day in an afterschool program and still be included in the treatment group. Meanwhile, non-participants may be enrolled in other afterschool activities and still be included in a control group without careful examination of their background characteristics. It is also rare that studies of afterschool programs consider students' prior attendance history in the program. Prior program attendance may influence the likelihood of current attendance, current performance on achievement outcomes, or both. In order to thoroughly understand the relationship between program attendance and achievement outcomes this study introduced the entropy balancing method together with the HLM analyses to account for three important issues:

1. The intensity of program attendance in a given year.
2. The consistency of attendance over time.
3. Background differences in treated and control populations may affect future program attendance and performance on achievement outcomes.

This methodology examined three years of sequential program treatment history and achievement outcomes for two cohorts of second grade students who initially did not attend the program. By confining the analyses to students who did not attend the program in second grade, a potential source for self-selection bias was removed. For math and English language arts outcomes we presented six models with treatment defined at increasing levels of program attendance. In addition, we identified four specific treatment effects that allowed for the examination of potential impacts regarding the consistency of attendance over time. This approach allowed for tracking the potential that afterschool treatment effects on achievement may require a dosage threshold of some combination or level of program attendance both within a year and across time. Analyses that only examine one of these dimensions may fail to identify important program effects. The interpretation of effects across these two dimensions also helped identify any lack of stability in the pseudo-samples we created. If, for example, positive significant effects are present at some attendance intensity level and then disappear at a higher intensity level this might suggest a problem with extreme weights or the influence of an un-measured confounder. Thus, examining the trends in the resulting effects reduces the likelihood of promoting a false finding as a true program effect.

In addition, we control for self-selection bias by directly balancing the defined treatment and control groups in each grade level and for each analyses on a set of prior observed covariates. We measured prior covariates at both the student and school levels, and included prior treatment indicators. We examined the weights produced by this balancing procedure to be confident that the treatment and control groups exhibited adequate common support. Despite these steps, we recognize that some un-measured confounder of self-selection and achievement might still have influenced our findings. We examined the area under ROC curves to determine that our findings may be vulnerable to such a potential confounder primarily in third grade.

Implication of Results

Implications from this study highlight the finding that simple indicators of program participation may be inadequate to capture program effects fully. If our study simply defined treatment as any program attendance, we would have found no evidence of program effectiveness with regard to math achievement. It would be useful for future researchers with access to longitudinal program attendance data conduct similar analyses that take into account the intensity and duration of program attendance.

More importantly, study results suggest that recruitment efforts as early as second grade, and retention efforts to keep these students in the program may lead to later math achievement. However, there appears to be room for LA's BEST to improve on early program recruitment and student retainment in their program. Of the 12,516 students in the second grade math sample, 1,813 moved into LA's BEST treatment

in third grade, yet only 523 of these students maintained an average of three days per week. Fewer still (n = 190) were able to maintain an average of four days per week during the three consecutive years. For a program to have impact on students' achievement, the students need to receive sufficient dosage. Supporting previous studies (Frankel and Daley, 2007), this study also suggests that regular afterschool program attendance (three days per week) for consecutive years may be necessary to reap program benefits on achievement outcomes. Thus, LA's BEST can improve its effectiveness by finding techniques to encourage all students to participate at this level.

Conclusion

This study set out to fill a research gap by demonstrating the use of rigorous methodology to study the effects of "dosage" (intensity of afterschool attendance) on students' academic outcomes. Entropy balancing can be an efficient tool to reduce the challenges on selection bias in afterschool studies. More concretely, this study tracked approximately 34,000 students for four years. It was found that students who attended regularly showed achievement growth in math. Results also suggested that early program participation from second grade on may also lead to better math performance. This study confirmed previous studies in further emphasizing the importance of regular participation in afterschool programs in order to reap program benefits.

References

Bradshaw, C. P., Wassdorp, T. E., Goldweber, A., Johnson, S. L. (2013). Bullies, gangs, drugs, and school: Understanding the overlap and the role of ethnicity and urbanicity. *Journal of Youthg and Adolescence, 42*(2), 220–234. https://doi.org/10.1007/s10964-012-9863-7

Caliendo, M., Kopeinig, S. (2005). *Some practical guidance for the implementation of propensity score matching.* Institute for the Study of Labor, IZA.

California AfterSchool Network. (2007). California's funding landscape. Retrieved December 17, 2007, from http://www.afterschoolnetwork.org/as_landscape

Durlak, J. A., Weissberg, R. P., Pachan, M. A. (2010). A meta-analyses of afterschool programs that seek to promote personal social skills in children and adolescents. *American Journal of Community Psychology, 45*, 294–309. https://doi.org/10.1007/s10464-010-9300-6

Frankel, S., & Daley, G. (2007). *An evaluation of after school programs provided by Beyond the Bell's partner agencies*. Los Angeles, CA: Beyond the Bell Branch, LAUSD.

Goldschmidt, P., Huang, D., & Chinen, M. (2007). *The long-term effects of after-school programming on educational adjustment and juvenile crime: A study of the LA's BEST after-school program*. Los Angeles, CA: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Hainmueller, J. (2011). Entropy balancing for causal effects: A multivariate reweighting method to produce balanced samples in observational studies. *Political Analysis, 20*, 25–46. https://doi.org/10.1093/pan/mpr025

Hodgkinson, H. (2006). *The whole child in a fractured world*. Alexandria, VA: Commission on the Whole Child, convened by the Association for Supervision and Curriculum Development.

Hollister, R. (2003). *The growth in after-school programs and their impact.* Washington, DC: Brookings Institution.

Hong, G., & Raudenbush, S. W. (2008). Causal inference for time-varying instructional treatments. *Journal of Educational and Behavioral Statistics, 33*(3), 333–362. https://doi.org/10.3102/1076998607307355

Huang, D., Gribbons, B., Kim, K. S., Lee, C., & Baker, E. L. (2000). *A decade of results: The impact of the LA's BEST after school enrichment program on subsequent student achievement and performance*. Los Angeles, CA: UCLA Center for the Study of Evaluation.

Huber, M., Lechner, M., & Wunsch, C. (2010). *How to control for many covariates? Reliable estimators based on the propensity score*. Institute for the Study of Labor, IZA.

LA's BEST. (n.d.). *What we do*. Retrieved from http://www.lasbest.org/what

Lauer, P. A., Akiba, M., Wilkerson, S. B., Apthorp, H. S., Snow, D., & Martin-Glenn, M. L. (2006). Out-of-School-Time Programs: A Meta-Analysis of Effects for At-Risk Students. *Review of educational research, 76*(2), 275–313 https://doi.org/10.3102/00346543076002275.

Little, P. M. D., & Harris, E. (2003, July*). A review of out-of-school time program quasi-experimental and experimental evaluation results.* Out-of-School Time Evaluation Snapshot, 1.

Maynard, B. R., Peters, K. E., Vaughn, M.G., Sarteschi, C.M. (2013). Fidelity in afterschool program intervention research: A sysmetic review. *Research on Social Work Practice, 23* (6), 613–623. https://doi.org/10.1177/1049731513491150

McKinsey & Company. (2009). *The economic impact of the achievement gap in America's schools*. Washington, DC: Social Sector Office.

Munoz, M. A. (2002). *Outcome-based community-schools partnerships: The impact of the after-school programs on non-academic and academic indicators*. Retrieved from ERIC database. (ED468973)

No Child Left Behind Act of 2001, Pub. L. No. 107–110, 115 Stat. 1425 (2002). Retrieved December 17, 2007, from http://www.ed.gov/legislation/ESEA02/

Ponitz, C. C., McClelland, M. M., Matthews, J. S. & Morrison, F. J. (2009). A structured observation of behavioral self-regulation and its contribution to kindergarten outcomes. *Developmental Psychology, 45* (3), 605–619 https://doi.org/10.1037/a0015365

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Thousand Oaks, CA: Sage Publications.

Robins, J., Herna'n, M., & Brumback, B. (2000). Marginal structural models and causal inference in epidemiology. *Epidemiology, 11*, 550–560. https://doi.org/10.1097/00001648-200009000-00011

Rogosa, D., Brandt, D., & Zimowski, M. (1982). A growth curve approach to the measurement of change. *Psychological Bulletin, 92*(3), 726–748. https://doi.org/10.1037/0033-2909.92.3.726

Schaps, E. (2006). *Educating the Whole Child.* Alexandria, VA: Commission on the Whole Child, convened by the Association for Supervision and Curriculum Development.

Scott-Little, C., Hamann, M. S., & Jurs, S. G. (2002). Evaluations of after-school programs: A meta-evaluation of methodologies and narrative summary findings. *American Journal of Evaluation, 23*(4), 387–419. https://doi.org/10.1177/109821400202300403

Sekhon, J. S. (2009). Opiates for the matches: Matching methods for causal inference. *Annual Review of Political Science, 12*, 487–508. https://doi.org/10.1146/annurev.polisci.11.060606.135444

Snijders, T. A. B., & Bosker, R. J. (1999). *Multilevel analysis: An introduction to basic and advanced multilevel modeling.* London, England: Sage.

TASC: The After School Corporation, (2005). *Quality, scale, and effectiveness in afterschool programs, summary of 2004 Policy Studies Associates' evaluations*. New York: Author.

U.S. Department of Education. (2003). Scientifically based evaluation methods. *Federal Register 68*(213), pp. 62445–62447.

Vanderhaar, J., & Muñoz, M. A. (2006). *Educating at-risk African American males: Formative and summative evaluation of the Street Academy Program*. Retrieved from ERIC database. (ED495958)

Welsh, M. E., Russell, C. A., Williams, I., Reisner, E. R., & White, R. N. (2002). *Promoting learning and school attendance through after-school programs: Student-level changes in educational performance across TASC's first three years.* Washington, DC: Policy Studies Associates, Inc.