Making Progress:
Longitudinal Outcomes Evaluation of the Say Yes to Education Program Kindergarten-5th Grade

Herbert M. Turner, III, Ph.D.
Jason Schoeneberger, M.A.
Tracey Hartmann, Ph.D.
Eva Gold, Ph.D.

MAY 2007

RESEARCH for ACTION

ANALYTICA
Research Analytics for the Social

__ANALYTICA__ is a for-profit, minority-owned and operated company founded on October 15, 2004 to provide high-quality research analytics to organizations in the social, behavioral, educational, and heath-care sectors. ANALYTICA is certified under U.S. Small Business Administration (SBA) guidelines. The founder and President of ANALYTICA is Dr. Herbert M. Turner, III.

__RESEARCH FOR ACTION__ is a Philadelphia-based, non-profit organization engaged in education research and evaluation, founded in 1992. Through research and action, Research for Action seeks to improve the education opportunities and outcomes of urban youth by strengthening public schools and enriching the civic and community dialogue about public education. We share our research with educators, parent and community leaders, students, and policy makers with the goals of building a shared critique of educational inequality and strategizing about school reform that is socially just.

ANALYTICA, Inc.
35 Goldfinch Circle,.
Phoenixville, PA 19460
ph: 215-808-8880
fx: 610-933-1005

Research for Action
3701 Chestnut Street
Philadelphia, PA 19104
ph: 215.823.2500
fx: 215.823.2510
www.researchforaction.org.

Making Progress:
Longitudinal Outcomes Evaluation of Say Yes to Education, Kindergarten-5th Grade

Prepared by
ANALYTICA, Inc.
and
Research for Action

Herbert M. Turner, III
herb@analytica-inc.com
ANALYTICA

Jason Schoeneberger
jason@analytica-inc.com
ANALYTICA

Tracey Hartmann
thartmann@researchforaction.org
RESEARCH FOR ACTION

Eva Gold
egold@researchforaction.org
RESEARCH FOR ACTION

Table of Contents

I. Introduction and Summary of Findings

Say Yes to Education (SYTE) is a scholarship guarantee program that pledges to young children from disadvantaged backgrounds a fully paid, post-secondary education along with academic and social supports that follow children and their families throughout their elementary and high school careers. Since its inception in 1987, the SYTE program has “adopted” cohorts of students in Philadelphia, Hartford, CT, Cambridge, MA, and New York City.

Research for Action (RFA) has conducted two evaluations of the Philadelphia SYTE chapter. This report shares the results of the second evaluation conducted in partnership with ANALYTICA, Inc. The second evaluation began in 2006 when the cohort was in fifth grade and focused on student outcomes. The first RFA evaluation of the Philadelphia SYTE program took place in 2003-2004, when the current cohort was in third grade. RFA’s first evaluation included qualitative research as well as an outcomes analysis. The qualitative research focused on understanding program processes and parent perspectives on the program. The outcomes analysis compared SYTE students to a similarly matched group at the end of third grade.

The qualitative research in the first evaluation found that a strength of the SYTE program was its highly relational orientation which made a priority of creating trust between staff and children, staff and families, parents and children, as well as among parents and among the children. SYTE staff developed services for families in response to needs that emerged over time and created the conditions that would help families support the SYTE students graduating from high school and being able to take advantage of the post-secondary scholarship. The first evaluation found that:

- SYTE parents had developed a high level of trust in the SYTE program and particularly its program manager;
- SYTE had created a sense of community among SYTE families so that families were beginning to support each other;
- SYTE was also creating a peer group of students that supported each other and had similar expectations for academic achievement; and,
- SYTE had been instrumental in changing and increasing parents’ involvement with their children’s education.

The outcomes analysis found that SYTE students as a group were performing better than other third graders in their school and the School District of Philadelphia (SDP). However, when compared to a similar group matched on reading levels at the end of second grade, the analysis was not able to detect a significant impact of the SYTE program on students' academic and behavioral outcomes in third grade. There was some evidence that suggested SYTE may have helped students who were weaker readers perform better than a comparison group in math.

A rigorous impact analysis of the Philadelphia SYTE program is a challenging task. The program has been underway for some time making random assignment impossible. In addition, there was a degree of self-selection into the SYTE program. Families who were initially chosen at random to receive the SYTE award chose not to participate and others contacted SYTE to be put on a waiting list for entry into the program. The SYTE program also serves a relatively small number of students, 45, giving any outcomes analysis limited statistical power to detect an effect. A power analysis shows that SYTE would need a sample of 800 (400 students in each group) to detect a small program effect. RFA's first outcomes analysis also had several other limitations. First, the comparison group was pulled from a limited pool of students--a cohort of students in the same elementary school, one year ahead of the SYTE cohort. This made it difficult to get a strong match for SYTE children. In addition, the first outcomes analysis used one covariate, reading level, to match SYTE students with a comparison group. The SDP assessment of reading levels changed between the first grade year of the matched group and the first grade year of the SYTE students and therefore first grade reading level marks were not comparable. Matching was done through the use of second grade reading levels. SYTE students had already begun to receive significant reading supports by the end of second grade.

SYTE asked RFA to improve upon this analysis looking at SYTE student outcomes in fifth grade. RFA sub-contracted with ANALYTICA, Inc. to design and conduct a study that could address some of the challenges of the first SYTE outcomes evaluation. This report presents the results of that post-hoc outcomes analysis. It assesses the impact of the SYTE program on the academic and school-based behavioral outcomes for Philadelphia SYTE students over their first six years in the program, from the time the SYTE students entered the program in kindergarten[1] (fall 2000) to when they completed their sixth year in the program in fifth grade (2005-06).[2]

The evaluation was designed to address the following three questions:

1. Are SYTE students "on-track" to graduate from high school and attend post-secondary institutions?
2. What is the discernible impact of the Say Yes program on participant performance on standardized achievement tests, promotion rates, course grades, attendance, and behavior marks annually and if there is a discernible impact, does it vary by gender?
3. If a discernible impact of the Say Yes program exists, does the impact vary over time (from the first year of the program to the fifth)?

The School District of Philadelphia provided the data necessary to answer these questions. The first question was addressed through descriptive data in the students fifth grade year. The second and third questions were addressed through cross-sectional comparisons of the SYTE students and a comparison group. The comparison group was developed through propensity score matching. The following chapters of this report provide a detailed description of the matching process. Given the small study sample size due to the small number of SYTE students used to create the matched comparison group, it was difficult to detect small and medium size program impacts. Therefore, in this study impacts were detected by examining "educationally meaningful" effect sizes. Effect sizes are differences between the mean of the comparison group and the mean of the treatment group divided by the pooled standard deviation of both groups. They are expressed in standard deviation units. Effect sizes that are larger than .25 are considered educationally meaningful (Lipsey & Wilson, 2001).

The analysis improves upon the outcomes analysis of the previous report in several respects. First, it matches SYTE children with a comparison group in kindergarten, before the SYTE program was fully underway. Second, it drew upon the entire SDP kindergarten dataset, with a sample size of approximately 23,000, to create a matched group. Seventeen covariates, measured during the students' kindergarten

[1] Two SYTE students did not enter the program in kindergarten. One entered in 1st grade and the other in 2nd grade.
[2] SYTE students who are "on-grade level" should be in fifth grade at the end of their sixth year in the program.

year, were used to create the match. Several outcome variables were used including two standardized tests (PSSA & Terra Nova's), attendance, grades and suspensions. In addition, the analysis looks at outcomes over time, comparing SYTE to a matched group at the end of each school year from 1st through 5th grade.

Data analysis as well as interpretation of the findings was informed by the previous research conducted by RFA. Therefore, this report is a joint report of ANALYTICA and Research for Action. This analysis represents the most comprehensive and rigorous evaluation of SYTE's work to date.

The second outcomes analysis began when the SYTE students were in fifth grade (2005-2006). Several important changes in the program had taken place since the first evaluation. In the 2004-2005 school years, the program lost its office and resource room in the neighborhood school, although it continued to use the school as home base for an after-school program and their summer Freedom School program. Second, the bulk of the Philadelphia SYTE students moved in the 2005-2006 school year from their neighborhood elementary school to middle schools around the city. The largest group was guided by SYTE to enroll in KIPP Academy Charter School. Others attended magnet schools and a few remained at the original neighborhood elementary school or other neighborhood schools. Third, beginning in September 2004, SYTE also went through a dramatic expansion, adopting five cohorts of children in Harlem, NYC. In 2006 SYTE experienced a transition leadership. With the dramatic expansion in size and new leadership, the program began to reflect on its core program elements as well as its cost-effectiveness. And, after 20 years of operation, the program was beginning to look for lessons learned which could inform public policy. Additional research was planned including a randomized controlled trial in a new city. The current outcomes analysis took on new importance within this context. Therefore, the research design was subject to peer review by American Institute for Research (AIR) before it was conducted.

Summary of Findings

The majority (85%) of SYTE students does not exhibit any of the risk factors of dropping out and appears to be "on-track" according to several indicators, to graduate from high school. Their academic performance on standardized tests however, remains worrisome. SYTE students were also performing better than a comparison group on a number of academic and behavioral outcome areas and across some years.[3]

Descriptive data is used to answer the question of whether SYTE students are "on-track" to graduate from high school and attend college. Our analysis of SYTE students being on-track to graduate from high school draws on the work of Jerald (2006) and Neild & Balfanz (2006) who looked at predictors of high school drop-outs. They point out several early warning signs in 6th, 8th or 9th grades which indicate that a student is not likely to graduate from high school. These warning signs include: being below grade level, having transferred among multiple elementary schools, a drop in academic performance and behavior after a transition to middle school, reading and math scores significantly below grade level, and 80% or lower attendance in 6th grade. Although it is still too early to tell definitively whether SYTE students will graduate high school, an examination of their fifth grade data in these important dimensions suggests that most of the SYTE students are "on-track" to graduate. A small group of 5-6 students are exhibiting some of the risk factors for dropping out.

- 39 SYTE students are on grade level, 6 students are one grade level behind. No SYTE students are more than one year behind.
- SYTE students have not transferred frequently between elementary schools. The majority remained at their local elementary school through 5th grade because the SYTE program was based at the school and encouraged families to remain there.
- After the transition to middle school, SYTE students' academic performance has not declined. Some evidence, in fact, suggests that students who transferred to KIPP charter school have made gains.
- On average, SYTE students were attending school 94% of the time, missing an average of 12 school days each year. Only 3 students missed 20% or more days of the school in their fifth grade year. Contributing to the high average number of excused absences are health related issues including five SYTE students who are frequently hospitalized for asthma. In reviewing this finding, SYTE staff added that a small number of SYTE families were truant in early years and this pattern disappeared with SYTE

[3] We use the term "impact" in the quasi-experiment, rather than the experimental, sense because there are always reservations with a propensity score matching design in attributing all of the observed group differences to the intervention, which in this study is SYTE.

monitoring of the problem and with greater parent-child participation in SYTE programs. Our data showed that SYTE families averaged 50 absences in kindergarten and this number dropped to 13 in first grade.[4]

- Few SYTE students were suspended in their fifth grade year. Four students were suspended one time and two students were suspended four times. The SYTE group has averaged less than one suspension each year although the number has also increased each year.
- SYTE students' performance on standardized tests is, however, worrisome. Only 13 of SYTE students scored proficient or advanced (on grade level) on the fifth grade state standardized test (PSSA) in math. Only 11 scored proficient in reading on the same test. Sixteen scored advanced or proficient in writing. Therefore, over half of the students are performing below grade level in these core subjects. Eighteen students were in the lowest category, below basic in math and 23 students were below basic in reading. None were below basic in writing but 20 were at the basic level.

While fifth grade behavioral and academic performance has not been found to be predictive of graduating or dropping out, similar outcomes in 6th, 8th or 9th grade are highly predictive of graduating or dropping out. Therefore, on-going monitoring of students outcomes is important for SYTE. Descriptive analysis of sixth grade outcomes could build upon this analysis and be even more telling in determining whether students are "on-track" to graduate.

The outcomes analysis was interested in not only whether SYTE students were "on-track" but whether they were performing better than non-SYTE peers as a result of all the supports they receive from SYTE. Using the "educationally meaningful" definition of impact, SYTE students were performing better than the comparison group with respect to cohort retention, promotion rates, unexcused absences, suspensions, and grades, and standardized test scores, although not for each year in the analysis. SYTE students were also more likely to receive support services for both special education and giftedness. Each outcome area will be discussed below:

[4] The comparison group also had a dramatic decrease in absences from kindergarten to first grade going from an average 49 days absent to an average of 17 days absent.

SYTE students demonstrated less mobility, were more likely to receive needed support services, and had greater parental cooperation with the school than the comparison group.

- SYTE students were more stable and less likely to leave their neighborhood school and the SDP than the comparison group. Our earlier research showed that families were influenced to keep their children at their original neighborhood elementary school even when they had moved out of the neighborhood. While it was not a requirement of the SYTE program to remain at the neighborhood elementary school, the SYTE program had an office and a resource room in the building and SYTE staff were on-site to provide additional supports and a safe space for students and parents during the school day. Many parents chose to keep their children at the school because of SYTE. Parents reported having serious concerns about the climate of the school but felt that the SYTE program buffered and protected their children from many of the challenges of the school.
- In addition, more SYTE students were receiving needed special education services for disabilities as well as giftedness and these were identified earlier than comparison group students. SYTE staff arranged monthly team meetings with classroom teachers and the principal in the early grades. These meetings were helpful in identifying children who needed extra supports.
- SYTE students have fewer unexcused absences and more excused absences than the comparison group in second grade, fourth grade, and fifth grade. This means that SYTE parents were more likely than parents of comparison group students to notify the school of the reason for their child's absence and suggests that parents are more engaged or cooperative with the school. However, SYTE overall number of absences did not differ from the comparison group.

Together, these findings suggest that SYTE has created a more stable and supportive context for learning than the one experienced by the comparison group. SYTE has created this context by providing school-based services in the early years, advocacy for students within the school and making extensive efforts to engage parents.

Overall attendance for SYTE students did not differ from the comparison group. However, SYTE students were less likely to be suspended than the comparison group in early elementary school.

- SYTE students attended the same amount of school as the comparison group. Both groups missed an average of 12 school days each year.
- SYTE had fewer suspensions than the comparison group in second and third grades. No SYTE students were suspended in first and second grade. According to SYTE staff, in response to this finding, the presence of the SYTE program at the school allowed teachers of students acting inappropriately to send students to the SYTE resource room rather than suspend them. In the SYTE room, students would complete their class work and receive additional academic supports which prevented them from falling behind in their lessons.

SYTE students were more likely to stay on grade level (ie., be promoted) than the comparison group in fourth and fifth grades.

- In 4th grade 100% of SYTE students were promoted while only 79% of the comparison group was promoted.
- In 5th grade, 98% of SYTE students were promoted while only 88% of the comparison group was promoted.

Again, the intense monitoring of, and advocacy for, SYTE students may have contributed to these findings. SYTE put in place special supports to help children avoid retention including providing an approved academically-rigorous summer program for students who were retained, threatened with retention or reading below grade level. This program ran for six weeks, full day, with small class sizes, certified teachers, and additional assistants to facilitate instruction.

SYTE performed better than the comparison group on some academic outcomes in some years. The effects were strongest for the entire group in early elementary school and evident again at the end of fifth grade. An impact was evident for girls across all years of the analysis, particularly in their science test scores.

- SYTE students had higher course grades in math, science, reading, and writing than the comparison group in first and second grade. This achievement can be attributed, in part, to the multiple supports for literacy and numeracy learning SYTE provided in the classroom as well as in after-school time, during the summer and through special math workshops for parents.
- SYTE students outperformed the comparison group on Terra Nova language arts, reading, math, and science exams in third grade. No differences between the SYTE group and the comparison group on the Terra Nova were observed in fourth or fifth grade. It is important to note that both the fourth and fifth grade Terra Nova exams were testing fourth grade learning. The SDP changed the timing of the Terra Nova exams to make the Terra Nova a diagnostic exam. Therefore, the fourth grade Terra Nova exam was given in the spring of the fourth grade year, the fifth grade Terra Nova was given in the fall of the fifth grade year. The fourth grade year was a particularly difficult one for SYTE students because they experienced the closing of the SYTE resource room and a long-term substitute teacher for one fourth grade class.
- SYTE girls outperformed the comparison group on the Terra Nova exams in each area, each year of the analysis with two exceptions (2nd grade science, 5th grade math). The difference from the comparison group was educationally meaningful in the following years and areas:
    - Second grade: math, language arts, spelling and word recognition
    - Third grade: science
    - Fourth grade: math and science
    - Fifth grade: reading and science
- SYTE boys outperformed comparison boys on the Terra Nova exams twice:
    - Second grade: spelling
    - Fifth grade: Math

- However, SYTE boys were also outperformed by the comparison boys in several years and areas:
    - Second grade: math and science
    - Fifth grade: reading and science
- While SYTE students performance on the state standardized PSSA was worrisome, they performed better than a comparison group on the PSSA math and writing tests in fifth grade. The SYTE average was brought up by the KIPP student scores. Unlike the results for the Terra Nova exams, there were no gender differences.

The differences in fifth grade Terra Nova and PSSA results could be explained by the timing of the exams. The Terra Nova was administered in the fall of the 5th grade year and thus reflects 4th grade learning. The PSSA is given in the spring of the fifth grade year and would reflect a full year of learning from 5th grade when many students had transferred out of their neighborhood elementary school to magnet schools or KIPP academy charter school. PSSA & Terra Nova scores are highly correlated and thus some comparisons between the two tests can be made. The PSSA results suggest that SYTE students made gains in their fifth grade year or at least, did not lose ground as a result of the transition. However, a comparison between fifth and sixth grade Terra Nova scores is required to confirm this progress.

While this analysis improves on previous research, it encountered new challenges and complications which should be kept in mind when reading this report. The lessons learned from the challenges of this research should inform future research. Subsequent chapters will describe these challenges in more detail as well as the strategies used to address them.

- The kindergarten data file obtained from the SDP was missing as much as 50% of the grade information needed for creating the propensity score match and thus data was imputed.[5] Future research could create a match in first grade where more data is available.
- In addition, the comparison group suffered significant attrition over the years of the analysis. Future studies should have the resources to track and obtain records for comparison group students who leave the school district. Future

[5] Some of the kindergarten grades are optional for kindergarten teachers to submit.

research could also create a larger comparison group to address the problem of attrition and statistical power. We attempted to identify additional matches in the first round of the analysis but their similarity to the SYTE on the matching variables decreased while not increasing the statistical power significantly.

- The use of zip code as a proxy for neighborhood characteristics significantly reduced the pool of students available for matching and made it difficult to find more than one good match for each SYTE student. We recommend that future research find other variables besides zip code to account for neighborhood characteristics.
- It is conceivable that SYTE could impact a number of psycho-social outcomes such as self-esteem, locus of control, and attitudes towards learning—none of which were outcomes in this evaluation.

Nonetheless, this analysis offers one of the most rigorous analyses of SYTE work to date and provides the best empirical estimate of the SYTE program effect. In addition, it has created an extensive database on SYTE students which future research could easily build upon to continue tracking student progress. The analysis also raises an important question for the work of SYTE; what is the relative contribution of the school to student learning as compared to other ancillary academic supports provided by SYTE to student progress? Future research with larger sample sizes and students in multiple schools should explore this question.

In what follows, first we present the research design and background to orient the reader as to the methodological strengths and challenges of use of propensity score matching with 2000-01 SYTE kindergarten cohort. Our intention is not an exhaustive expose, but to introduce the reader to the fundamentals of the technique and conditions that warrant its use. For a more extensive discussion of propensity-score matching, the reader is referred to Victor (2007). Second, we extend this discussion more broadly to methods. Specifically, we described how the evaluation design generates the data to address the research questions, and how that data was collected and analyzed. Third, we present the results. As we will show, the research design allows the reader to consider the impact of the SYTE program on student outcomes for a particular program year and across program years. We close with methodological recommendations for designing future rigorous evaluations of the SYTE.

II. Research Design: Rationale for Propensity Score Matching

The evaluation used propensity score matching to create a matched group of School District of Philadelphia kindergarteners to which the SYTE students could be fairly compared over time. Why was propensity score matching, rather than random assignment, used to equate the SYTE and comparison kindergartners? At the time of the evaluation design, SYTE participants were in their sixth year of the program and, more importantly, had self-selected into the program[6] when it was launched. Under these circumstances, the best available design option is to control for as many observable characteristics (at the student, neighborhood, and school level) as possible that are theorized to be correlated with the outcomes on which the SYTE and comparison groups are compared (Dehejia, Wahba, 1999; and Luellen, Shadish, & Clark, 2005).

Propensity score matching is a statistical procedure designed to balance groups on observable characteristics that can be measured validly and reliably (Rosenbaum & Rubin, 1983). Propensity scores are the estimated probability that a program participant is assigned to an intervention based on observable variables (Pasta, 2000). Essentially, the predicted probability is obtained by conducting a logistic regression that predicts membership in the intervention group utilizing a vector of covariate predictors.

Theoretically, subjects with similar distributions across the covariates will have similar estimated propensity scores. A student receiving an intervention such as the SYTE program can be matched with a comparison student with a similar propensity score, generated by a logistic regression equation in which the observable covariates are the independent variables and intervention group status is the dependent variable. The result is a reduced-bias estimate of the intervention's impact on the outcome when the groups are compared using quantitative measures such as an effect size.

Conceptually, propensity score matching is the observational study analog of a randomized controlled trial, but is less effective in producing unbiased estimates as it can only balance the distribution of *observed* covariates, whereas randomization balances the distribution of *all* covariates, both observed and unobserved (Rosenbaum & Rubin, 1983). Whether propensity scores can approximate benchmark estimates of randomized controlled trials varies according to a number of factors including the type of

[6] While the program was offered to an entire cohort of head start children, some families did not respond to efforts by SYTE to contact them. Some families also opted out of the program. SYTE filled in open slots through lottery.

intervention (drop-out prevention or employment and training) and the conditions in which the study is conducted (e.g., the comparison group was drawn from within the evaluation itself rather than from a national dataset).

No doubt, propensity score matching has critics and proponents (Shadish, Luellen, and Clark, 2005). For example, Agodino and Dynarski (2004) found that for dropout prevention programs, there was a lack of consistent evidence that estimates from propensity scoring matching designs approximate those from experiments. However, Dehejia and Wahba (1999) found just the opposite result for labor training programs; that is, propensity score estimates of the impact of labor training programs are close to those of experiments.

Glazerman, Levy, and Myers (2003) conducted a systematic review to address the important question of "Do we know the conditions under which nonexperimental impact estimates are likely to replicate experimental impact estimates?" Encouragingly, these researchers identified some factors correlated with lower bias in program impact estimates. For example, bias in impact estimates was lower when the comparison group was drawn from within the evaluation itself rather than from a national dataset, when it was locally matched to the intervention group, and when it was itself drawn as a comparison group in an evaluation of a similar program or the same program in a different study site. As will be demonstrated in the methods sections, the conditions for lower bias in impact estimates as spelled at in the work by Glazerman and colleagues were present in this evaluation.

In addition, Victor & Boruch (2007) compared propensity score matching to twelve different types of statistical models through comprehensive simulations with large samples. They found that propensity score matching generated estimates with the least amount of bias, followed by Ordinary Least Squares Regression analysis (OLS).

Finally, it has been demonstrated empirically that creating matched groups using propensity scoring can reduce bias introduced by covariates by as much as 90% (Luellen, Shadish, & Clark, 2005; and Rosenbaum & Rubin, 1984). It must be acknowledged that there are always reservations regarding results generated by propensity score matching because it cannot equate groups on unobservable characteristics the way that random assignment can.

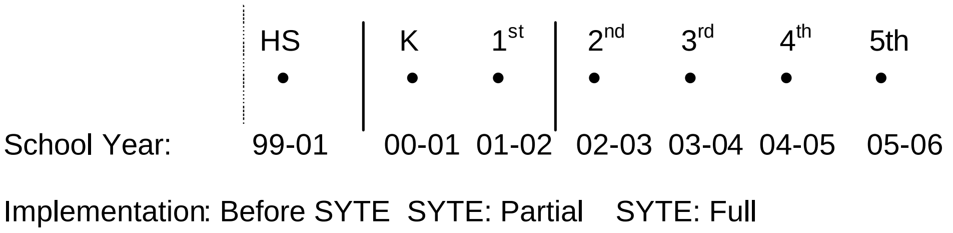
III. Methods

3.1 Baseline Covariates & Propensity Scoring

The SYTE program is designed to be multi-dimensional with academic and social supports for participants and their families. According to program staff, SYTE was implemented with moderate intensity during the first two years when students were in the kindergarten and first grade, respectively, but then with full intensity during the next four years—when students were on average in 2nd, 3rd, 4th, and 5th grades (see Figure 1). This variation in program implementation is noted as a consideration when interpreting impact estimates of the SYTE program.

To address two of the three questions for this evaluation, propensity score matching was, as stated earlier, used to create a fair comparison group for SYTE participants.

Figure 1. Implementation of the SYTE Program for the 2001 Cohort



HS K 1st 2nd 3rd 4th 5th

School Year: 99-01 00-01 01-02 02-03 03-04 04-05 05-06

Implementation: Before SYTE SYTE: Partial SYTE: Full

Note. HS = Head Start; K = Kindergarten; SYTE = Say Yes fo Education program;
Partial = Partial Implementation; and Full = Full Implementation

The propensity match procedure, implemented by ANALYTICA, used a logistic regression model with the following covariates (i.e., independent variables on the left-hand side of the logistic regression equation) observed during SYTE students and comparison students' kindergarten year:[7]

[7] The dependent variable was membership in Say Yes to Education.

- Ethnicity,
- Gender,
- Zip code (as proxy for neighborhood characteristics),
- School attended[8],
- Free or Reduced-Price Lunch (FRL) status,
- English Speaker of Other Language (ESOL) status,
- Grade level,
- Number of days absent from school,
- Number of days late to school, and
- Low-Performing school assigned to an educational management organization.[9]

The School District of Philadelphia included course grades for students in the data file for each of the years of interest in this evaluation for the following subjects:

- Math,
- Language Arts,
- Personal Growth,
- Work Habits,
- Physical Developments,
- Art & Music, Science, and
- Social Studies.

Alphabetical levels on report cards were converted to a numeric scale to represent an increasing level of ability and mastery. In reading and writing development, for example, Level D signaled a student with the ability to:

- Tracks words with eyes and not fingers,
- Uses pattern and language syntax to read with phrasing,
- Solves unfamiliar words with knowledge of letter-sound relationships.

Whereas Level H signaled a student with the ability to:

- Solve new words by using word analysis, then checking words against meaning,
- Reread to check and search,
- Discuss ideas from the story to indicate understanding.

[8] Note: this is not a multi-level model. A logistic regression model was used to estimate the propensity score.

[9] In 2000-01, the SDP was taken over by the State of Pennsylvania. As part of this takeover, 86 of the lowest performing schools were given over to outside managers, Educational Management Organizations, for reform. The school attended by the SYTE children was one of these 86 schools. To control for the unique school context, only schools that were part of this group of 86 were included in the propensity score matching process. This is a dichotomous indicator variable with "1" denoting a kindergartener's school designated by the School District of Philadelphia as 'low performing" and "0" denoting a kindergartner's feeder school as not "low performing."

These alphabetical marks were translated into a numerical scale to facilitate the calculation of means and standard deviations for conducting comparisons between SYTE participants and the comparison groups.

3.2 Dealing with Missing Data

Although missing data is inevitable when administrative records are used in a post-hoc evaluation, there was more missing data than expected on two behavioral variables ("days absent" and "days late") and report card variables in the dataset provided by the School District of Philadelphia. On average, 50% of the responses were missing on these variables. The data from students' kindergarten year—particularly the attendance and academic variables—were important for creating a reliable and valid propensity-matched comparison group. The academic variables were the only measure of achievement available for kindergarteners. If list wise deletion had been used, the kindergarten sample would have been reduced in half resulting in a substantial loss of the pool of kindergarteners from which to draw the comparison students rendering attempts to find matches based on propensity scoring less likely. To address this problem statistically, we used multiple imputation (MI). MI has statistical properties as good as can be hoped to achieve and is gaining currency among methodologists as a valid and reliable method for dealing with missing data, especially in post-hoc evaluation (Allison, 2001).[10]

[10] We emphasize the validity of the use of MI with post-hoc evaluations because the optimal solution for dealing with missing data is to design and implement a study in a way that minimizes missing data. Consistent with guidance provided by Allison (2001), we implemented MI using Statistical Analysis System (SAS) as follows:

1. Ran PROC univariate to generate descriptive statistics to determine the amount of missing data on all covariates;
2. Because our dependent variable is dichotomous, we transformed each of the variables (logit), prior to running MI, using the SAS Data Step and Array statements. We implemented PROC MI by including all the covariates with no missing data (then the two behavioral variables and report card variables) including Ethnicity, Gender, ESOL, FRL, along with the two behavioral and report card variables for which we are interested in imputing data. We implemented the procedure with five iterations resulting in five data sets with no missing data (i.e., the covariates that previously had missing data now had "imputed values" this values differed across the five data sets). Variables that were transformed on the logit scale were back transformed after the imputations were conducted;
3. We ran PROC Logistic, for each of the five datasets, with intervention group status as the depend-ent variable and the two behavioral variables and report card variables as the independent variables;
4. Finally, we ran PROC MIANALYZE to combine the estimates from each of the five datasets. Ultimately, use of MI restored the kindergarten sample to its original size. Equally important, and as will be shown in the results section, the restoration of sample did not alter the covariate's means values from what they were prior to implementation of MI, but as expected, the standard errors of the means were much lower after implementation of MI. The SAS code used to implement MI is presented in Appendix B.

3.3 Identifying Comparison Kindergarteners

The comparison group was drawn retrospectively or post-hoc from the total population of School District of Philadelphia children who were kindergartners in 2001. The first step in creating the comparison group was to generate the propensity for SYTE kindergartners (n=45) and comparison kindergartners (n=45). The logistic model used to generate the propensity score for kindergarteners, with no missing data on the covariates, took the following mathematical form:

$$(0.1) \qquad \mathrm{P}(\mathrm{SYTE}=1)=\frac{1}{\{1+\exp[-(\mathrm{B}_0+\mathbf{B}_1(\mathbf{X}))]\}} \text{ , where}$$

- **P(SYTE = 1)** is the probability that any student in the kindergarten sample would be assigned to the SYTE group (i.e., the propensity score);
- **B** is a vector of parameters for the 17 covariates in the model; and
- **X** represents the corresponding vector of 17 covariates enumerated earlier.

The predicted probability, or propensity score, serves as a single value that quantifies the *observable* covariate profile, with respect to assignment to SYTE, for every kindergartener in the sample.

After each kindergartener was assigned the propensity score (or probability of being assigned to the SYTE program), a “greedy” matching algorithm was applied, iteratively, to identify a kindergartner in the comparison pool with nearest propensity score to a kindergartener in the SYTE. For example, the first SYTE was selected for matching. All potential comparison students were randomly sorted in ascending order of their propensity scores to ensure that kindergartens with equivalent propensity scores have a random chance of being chosen as a comparison kindergartener. Once sorted, each potential comparison kindergartener’s propensity score was subtracted from the SYTE student’s propensity score to create a difference score. The kindergartener with the smallest difference value is then flagged as the match for the SYTE kindergartner. That SYTE kindergartner and the matched comparison kindergartner were removed from the file and the iteration for the next SYTE kindergartner was conducted. The iterative process continued for each SYTE kindergartner until a match for all SYTE kindergartners was identified resulting in the comparison group.

3.4 Comparisons of Group Outcomes

Figure 2 shows that using propensity scoring to match SYTE to comparison kindergartners results in two evaluation designs within one: 1) Cross sectional and 2) longitudinal. Cross-sectional or annual comparisons between the SYTE and comparison groups on academic and behavioral outcomes, as measured in School District of Philadelphia administrative records, were made at the end of *each* year of the program (e.g., 1st, 2nd, 3rd, and so on). In sum, one dimension of the evaluation design allowed for annual assessment of SYTE program impacts.

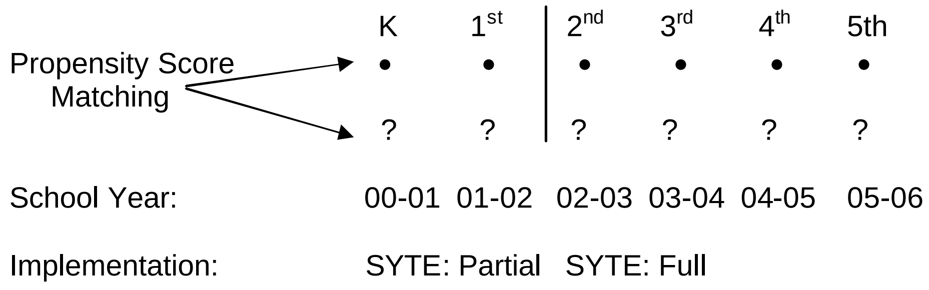
Specific outcomes that were compared included the following:

- Course Grades in first through fourth grade[11]
- Terra Nova math, reading and science exams in second through fifth grade
- Pennsylvania State Assessments (PSSA) for reading, math and writing in fifth grade
- Grade Promotion and Retention each year
- Day Absent, Excused and Unexcused absences each year
- Suspensions each year
- Special Education status.

<u>Cross-Sectional Comparisons</u>. To measure the impact of the Say Yes program on participants, the end-of-year outcomes for the Say Yes and comparison kindergarteners were compared using an effect size (and its standard error and confidence interval). As formula 1.2 shows, an effect size can be defined as the difference between the sample mean of the SYTE $\bar{X}_{SYTE}$ on an outcome and the sample mean of the comparison group ($\bar{X}_C$) on the same outcome.

[11] It is important to note the limitations in using grades in specific subjects as an outcome variable in this evaluation. Beyond kindergarten, patterns of assignment of alphabetical letters were neither consistent across each student's record nor across years. Thus, comparability of student grades across time, beyond kindergarten, had questionable validity. More important, the availability of course grades, in the administrative data file, for students in both the SYTE program and comparison group diminished substantially over time. For a complete accounting of administrative data that was requested and obtained from the School District of Philadelphia, see Appendix A.

Figure 2. Cross-Sectional and Longitudinal Comparisons of SYTE and Comparison Groups



Note. ● = Mean posttest outcome for SYTE Participants; ?= Mean posttest outcome for Comparison Group; K = Kindergarten; SYTE = Say Yes to Education; Partial = Partial Implementation; and Full = Full Implementation; and | = Marker interval denoting key implementation change in the Say Yes Program.

As formula (0.2) shows, this quantity is divided by the pooled standard deviation of the outcome for both groups:

$$d = \frac{\bar{X}_{SYTE} - \bar{X}_{C}}{\sqrt{\frac{(n_{SYTE}-1)s_{SYTE}^{2}+(n_{C}-1)s_{C}^{2}}{(n_{SYTE}+n_{C}-2)}}} \qquad (0.2)$$

An Effect Size (hereafter referred to as the d index or standardized mean difference), is an expression of the mean difference between two groups on an outcome that is expressed in standard deviation units, and is interpreted as the percentage of the standard deviation of the outcome.[12] In general, a standardized mean difference of 0.25 (or 25% of a full standard deviation) or larger is considered educationally substantive and important (Lipsey & Wilson, 2001). The standard error and confidence interval will be used to determine whether the observed difference between the SYTE Program participants and the comparison group, if one exists, is due to chance (i.e., is statistically significant). The formula used to compute the standard error is as follows:

[12] When the combined group sample sizes were 20 students or less, we multiplied the d index by a small sample correction factor of J = 1 - (3 / (4 * df - 1)) such that d*j results in a small sample corrected d index known as Hedges g.

(0.3) $$SE_d = \sqrt{\frac{1}{\mathrm{n_{SYTE}}+1} + \frac{1}{\mathrm{n_C}+1} + \frac{\mathrm{d}^2}{2*(\mathrm{n_{SYTE}}+\mathrm{n_C})}}$$

The 95% confidence interval for the standardized mean difference was computed using the following formula:

(0.4) $$95\% CI_d = d \pm t_{(a/2)} SE_d$$

The confidence interval conveys the same information as a test of statistical significance (when the interval crosses zero the non-zero effect size is interpreted as being due to chance) and has the additional interpretational advantage of conveying the range of effect sizes that would be observed, theoretically, in repeated samples of the same size (Kline, 2004).

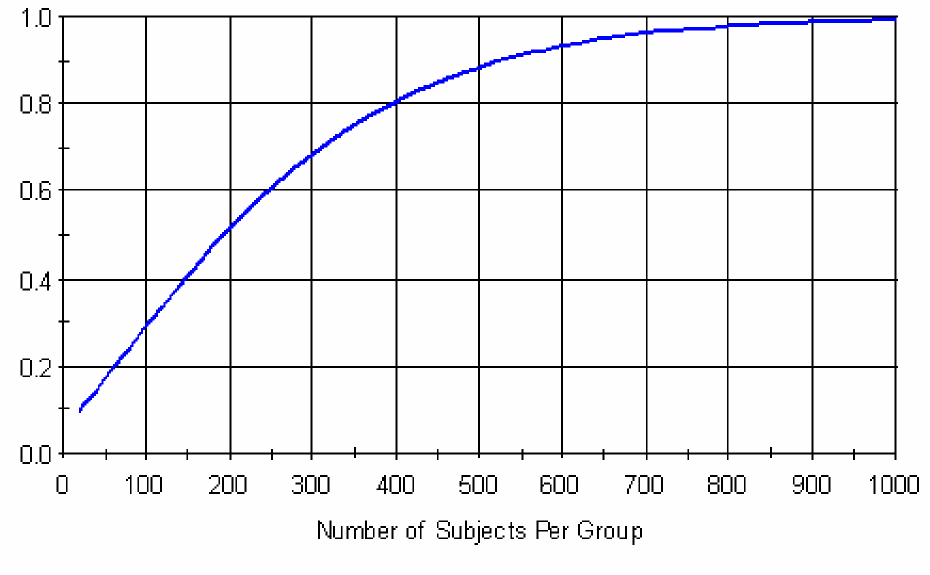
Longitudinal Comparisons. Two approaches were used to assess the trend in effect sizes computed from the end-of-program-year comparison of the two groups (i.e., annual comparisons between the SYTE group versus the comparison group). First, effect sizes were examined across years of the program (refer back to Figure 2). This allowed for a longitudinal comparison of the outcome trajectory for each group over the six years of the SYTE program. The second approach involved the use of a longitudinal regression modeling, to provide answers to whether any academic benefit accumulated for Say Yes students over time. The ability to model such growth is dependent on there being sufficient variance in changes in outcomes of interest across time (refer back to Figure 2).

3.5 Statistical Power

Statistical power of a research design can be defined as the ability of the design, when certain assumptions are met, to detect a statistically significant effect when one exists. Statistical power was constrained by the post-hoc nature of this evaluation and the small sample of 45 SYTE kindergarteners used to create the comparison group by finding a propensity score match in the comparison pool of kindergartners. Figure 3 was generated using Power and Precision software and illustrates the power of the two-group propensity scoring design as a function of the sample size (Borstein 1999).[13]

[13] The power is based on the following assumptions: $a = .05; d = .25$; and balanced groups.

Figure 3. Statistical Power for the SYTE Research Design



For this two-group (SYTE and comparison) design with 45 kindergartners per group, Figure 3 shows that the power for this design is below 0.20 meaning that less than 20% of designs with this sample size (n = 90) would detect a minimum effect size (*d*) of 0.20. In contrast, to achieve conventional power of 0.80, 400 subjects per group were needed. Thus, from the outset the statistical power of the research design was constrained by the small sample of SYTE kindergartners used to create the matched comparison sample.

We attempted to address this constraint by increasing statistical power through an unbalanced matching of SYTE kindergartners to comparison kindergartners. For example, we assumed an unbalanced match of 45 SYTE kindergartners to 360 comparison kindergarteners (i.e., a 1:8 SYTE to comparison group allocation ratio). The power analysis for this design increased power only slightly (0.22) and fell well short of the conventional target of 0.80. For this reason, we continued with the original plan of balanced allocation ratio or 1:1 propensity score match of SYTE and comparison kindergarteners.

The lack of statistical power was dealt with in two ways. First, we reported results using a Forrest Plot that displays the following:

1. Effect Size (d) - magnitude of the average difference between the two groups (e.g., effect size);

2. Confidence Interval – the variation in the effect size in repeated samples of the same size; and
3. P Value – statistical significance.

By reporting results using a Forrest Plot and not reporting p-values only, we make plain to the reader and differentiate between the magnitude of the intervention's effect (effect size) and whether this effect is due to chance (confidence interval and p-value). Second, we differentiate between an effect that is large enough to be educationally meaningful even if the size of the sample indicates the results may be due to chance (i.e., is not statistically significant). We do this by defining an *educationally meaningful* effect as an effect size that is 25% of one standard deviations or $d = 0.25$.

Calculation and presentation of effect sizes and their corresponding confidence intervals were chosen because this information can be used to serve two purposes: (1) to quickly realize whether the effect of interest was significant and (2) the practical significance of the effect. Effect sizes present the results in standard deviation units, allowing the reader to easily determine whether an effect, despite possibly being *non*-significant is meaningful in its own right. Confidence intervals that do *not* cross zero can be said to contain an effect size that is significant at the customary p < .05 level. Further, confidence intervals allow us to say that upon repeated experimentations, we can assume with 95% confidence that the effect size of interest would fall somewhere between the lower and upper bounds of the confidence interval. This information is important for understanding the inherent variability that may exist given the sample analyzed here is but one of many possible samples taken from the population of interest.

IV. Results

4.1 The SYTE 00-01 Cohort

We began by identifying SYTE participants as kindergarteners during the 2000-01 school year. Table 1 shows that three participants could not be identified in the district data during their Kindergarten year, but were subsequently identified in their first and second grade years (*n* = 2 in first grade, *n* = 1 in second grade). Table 1 also shows the number of Say Yes participants that were identified across time in the district report card data files.

Table 1. Say Yes Participants by Grade and School Year

	Grade Level							
School Year	*KG*	*1*	*2*	*3*	*4*	*5*	*Missing*	*Total*
2000-01	45							45
2001-02		47						47
2002-03		2	43				1	46
2003-04			7	37				44
2004-05				7	37			44
2005-06					6	21	18	45
Total	45	49	50	44	43	21		252

Initially, 45 SYTE participants were identified in their kindergarten year in the administrative data file. For the 2001-02 school year, two more participants were identified in Grade 1 for a total of 47, and an additional participant in the third year (2002-03) resulting in 48 SYTE participants. However, the overall total for 2002-03 decreased to 45 participants due to attrition of 3 SYTE participants. In addition, a number of SYTE (n =18) did not have grade information for the 2005-06 school year because no SDP report card was generated for those students during that year. At the recommendation of the SYTE program staff, the 18 SYTE participants attended KIPP Charter School in 2005-06 in which case their report card grades were not part of the School District of Philadelphia's (SDP) administrative records.

4.2 Missing Data and Multiple Imputation

For the 45 SYTE kindergartners (2000-01 school year) and 23,668 kindergarteners in the comparison pool, there were missing values on the course marks and attendance variables, as described in the methods section. Table 2 shows the percent of students with missing data on these key variables.

Table 2. Percent of Missing Data On Key Propensity-Match Predictors.

	Not Say Yes Participant			Say Yes Participant		
Variables	N	Missing	% Missing	N	Missing	% Missing
Days Absent	14637	8986	38.0	45	0	0.0
Days Late	14637	8986	38.0	45	0	0.0
Math	12491	11132	47.1	42	3	6.7
Language Arts	11630	11993	50.8	29	16	35.6
Personal Growth	12185	11438	48.4	29	16	35.6
Work Habits	12239	11384	48.2	23	22	48.9
Physical Development	12307	11316	47.9	26	19	42.2
Art & Music	11932	11691	49.5	5	40	88.9
Science	12122	11501	48.7	15	30	66.7
Social Studies	12420	11203	47.4	42	3	6.7

After implementing MI, there were valid values for each variable for all 23,668 kindergartners in the administrative data file; in other words, there was no missing data on any variable for any kindergartner in the data file. Table 3 shows the means, standard errors of the mean (SEM), and the minimum and maximum values for the Kindergarten variables before and after MI. It is important to note that the mean values for each variable remained virtually the same before and after the MI process as did the minimum and maximum values. As expected, the standard errors of the means were lower after the MI process, due to the larger sample sizes, than before the MI process was implemented.

Table 3. Descriptive Information for Kindergarten Variables Before and After MI.

	Before MI				After MI			
Variables	Mean	SEM	Min	Max	Mean	SEM	Min	Max
Days Absent	48.6	0.06	0	55	47.72	0.04	0	55
Days Late	6.25	0.06	0	55	7.09	0.04	0	55
Math	21.8	0.06	1	51	22.36	0.04	1	51
Language Arts	24.2	0.08	1	54	24.69	0.04	1	54
Personal Growth	9.5	0.03	1	21	9.58	0.02	1	21
Work Habits	7.18	0.03	1	15	7.31	0.02	1	15
Physical Development	17.6	0.04	1	48	17.78	0.02	1	48
Art & Music	2.29	0.01	1	6	2.31	0.00	1	6
Science	3.81	0.02	1	9	3.90	0.01	1	9
Social Studies	10.9	0.03	1	27	11.13	0.02	1	27

Note. The total sample size is 23, 668 kindergartners.

4.3 The Matched Comparison Group and Baseline Equivalence

The propensity score matching procedure was used to identify 45 kindergartners from the pool of 23,668 that were the best matches, defined as the minimum difference on the propensity score, for the 45 SYTE participants. The results of the propensity matching process for demographic variables are shown in Table 4 and Figure 4. Table 4 shows that the propensity match process created a relatively equivalent comparison group based on gender and ethnicity. Although statistical tests are suspect given the power constraints in this design, for completeness we examined the Chi-Square goodness of fit test and inferences to the population were consistent with the results in the sample, namely, there was no statistically significant relationship between the two demographic variables and group membership (SYTE and comparison): $c^2 = 0.18, p = .67$ and $c^2 = 3.1, p = .08$ for gender and ethnicity, respectively. However, there were three Latino SYTE kindergartners that did not have a match in the comparison group but matches were found for the remaining kindergarteners who were African American.[14] For the remaining demographic variables, Zip Code and Schools, the propensity score matching created a balanced comparison group. The results for the latter two variables are presented in Appendix B.

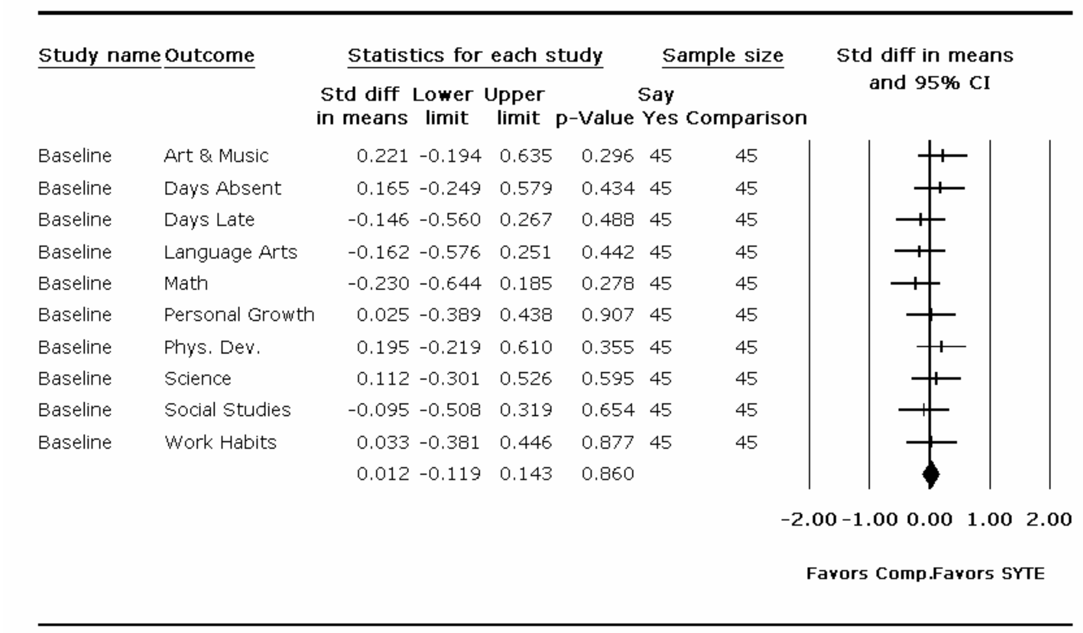
[14] It is interesting to note that we also stratified the comparison pool of kindergarteners by ethnicity and tried to create matched comparison groups accordingly. For example, we tried to create a comparison pools of White students but could not find matches. We attribute this to the use of zip code to control for neighborhood effects.

Table 4. Propensity Match Results for Demographic Variables

		Control		Say Yes		Total
Variable	Category	n	%	n	%	n
Gender	Female	23	51.1	21	46.7	44
	Male	22	48.9	24	53.3	46
	Total	45	100.0	45	100.0	90
Ethnicity	African American	45	100.0	42	93.3	87
	Latino	0	0.0	3	6.7	3
	Asian	0	0.0	0	0.0	0
	Other	0	0.0	0	0.0	0
	White	0	0.0	0	0.0	0
	Total	45	100.0	45	100.0	90

Figure 4 shows that SYTE and comparison group of kindergarteners were equivalent on Days Absent and Days Late; on the report card marks for core academic subjects of reading, math, science, and language arts; and on other report card marks such as personal growth. However, there was a borderline "educationally meaningful difference" that favored the comparison kindergartens in math (d = -0.23, 95%CI = -0.64 to 0.19) and the SYTE kindergartners in Art & Music (d = 0.22, 95%CI=-0.19 to 0.63).

Figure 4. Baseline Comparison of SYTE and Comparison Groups on Behavioral Measures and Report Card Marks



Study name	Outcome	Statistics for each study				Sample size	
		Std diff in means	Lower limit	Upper limit	p-Value	Say Yes	Comparison
Baseline	Art & Music	0.221	-0.194	0.635	0.296	45	45
Baseline	Days Absent	0.165	-0.249	0.579	0.434	45	45
Baseline	Days Late	-0.146	-0.560	0.267	0.488	45	45
Baseline	Language Arts	-0.162	-0.576	0.251	0.442	45	45
Baseline	Math	-0.230	-0.644	0.185	0.278	45	45
Baseline	Personal Growth	0.025	-0.389	0.438	0.907	45	45
Baseline	Phys. Dev.	0.195	-0.219	0.610	0.355	45	45
Baseline	Science	0.112	-0.301	0.526	0.595	45	45
Baseline	Social Studies	-0.095	-0.508	0.319	0.654	45	45
Baseline	Work Habits	0.033	-0.381	0.446	0.877	45	45
		0.012	-0.119	0.143	0.860		

Although 45 SYTE participants were identified as kindergarteners at baseline, a total of 48 Say Yes participants were ultimately identified across the entire longitudinal administrative data file:

- The majority (*n* = 45) of SYTE participants were flagged as kindergarteners in 2000-01;
- An additional two students as 1st graders in 2001-02, and one student as a 2nd grader in 2002-03;
- Propensity-based comparison students were also identified for these three additional students, however, the limitations in data provided by SDP required a different logistic model using the limited number of covariates available for the school years corresponding to when the three additional SYTE participants entered the program.
- These logistic models used to identify matches for the three additional SYTE participants used the following covariates to generate the propensity scores:
    - Ethnicity,
    - Gender,
    - Grade level,
    - LEP status,
    - LEP level,
    - Disability classification,
    - Number of suspensions,
    - Number of excused absences,
    - Number of unexcused absences,
    - Number of ‘other’ absences, and
    - Number of days enrolled.

Furthermore, because the SYTE students did not appear in the longitudinal administrative data file until later on, the comparison student pool was comprised of only those students that did not appear in the data file until the same year as the SYTE students of interest. Results from these iterative propensity matching routines generated results similar to those just reported.

4.4 Outcomes: Attrition and Grade Retention in SYTE and Comparison Groups

Before presenting outcome data on the impact of the SYTE program on behavioral and academic outcomes, we present a grade-level comparison between the SYTE kindergarteners and the comparison group kindergartners. **This comparison shows that the comparison cohort exhibited greater attrition and higher grade retention rates than the SYTE cohort.**

Table 5 displays the number of SYTE participants and the comparison group students in each grade from 2000-01 to 2005-06. In the first year, both the SYTE and comparison group comprised 45 students. In 2001-02, the comparison group was comprised of 36 students who were in 1st grade; one student was still in Kindergarten; two students had missing data so grade level could not be determined; and six students were no longer in the administrative data file. In other words, eight students dropped out of the comparison group for unknown reasons.[15]

During the same period, but in contrast, the SYTE group still comprised of 45 students all of whom moved to kindergarten, and two additional students were added to the cohort. Beyond this period, **the composition of the SYTE group was more stable than that of the comparison group.** The reader will notice that in 2005-06 there were 18 SYTE participants with missing data on grade level—these SYTE participants moved to KIPP charter school and, therefore, were tracked in the school's, rather than the district's, administrative record system.

Quantifying the level of retention exhibited by SYTE and comparison group was complicated by the level of attrition from the School District of Philadelphia in the comparison group. Further complicating matters was with the missing information on grade level for some SYTE participants in later years.

[15] There are a number of possible explanations for why comparison students were no longer in the database, none of which could be verified empirically, including moving out of the district or attending a charter school. ANALYTICA initially matched SYTE students with multiple students from the comparison group—a strategy that addresses the issue of attrition in the comparison group. However, this compromised the baseline similarity of the two groups. The results of the analysis based on a larger comparison group are available by contacting the authors of this study.

Table 5. Longitudinal Membership in Say Yes & Propensity-Based Control Groups

	Grade															
	KG		1		2		3		4		5		6		Missing	
School Year	Cntrl	Say Yes	Cntrl	Say Yes	Cntrl	Say Yes	Cntrl	Say Yes	Cntrl	Say Yes	Cntrl	Say Yes	Cntrl	Say Yes	Cntrl	Say Yes
2000-01	45	45														
2001-02	1		36	47											2	
2002-03			2	2	29	43										1
2003-04					5	7	20	37	1						4	
2004-05							8	7	16	37					3	
2005-06									7	6	14	21	1		3	18
Total	46	45	38	49	34	50	28	44	24	43	14	21	1	0	12	19

Table 6. Longitudinal Retention Rates for Say Yes & Comparison Groups

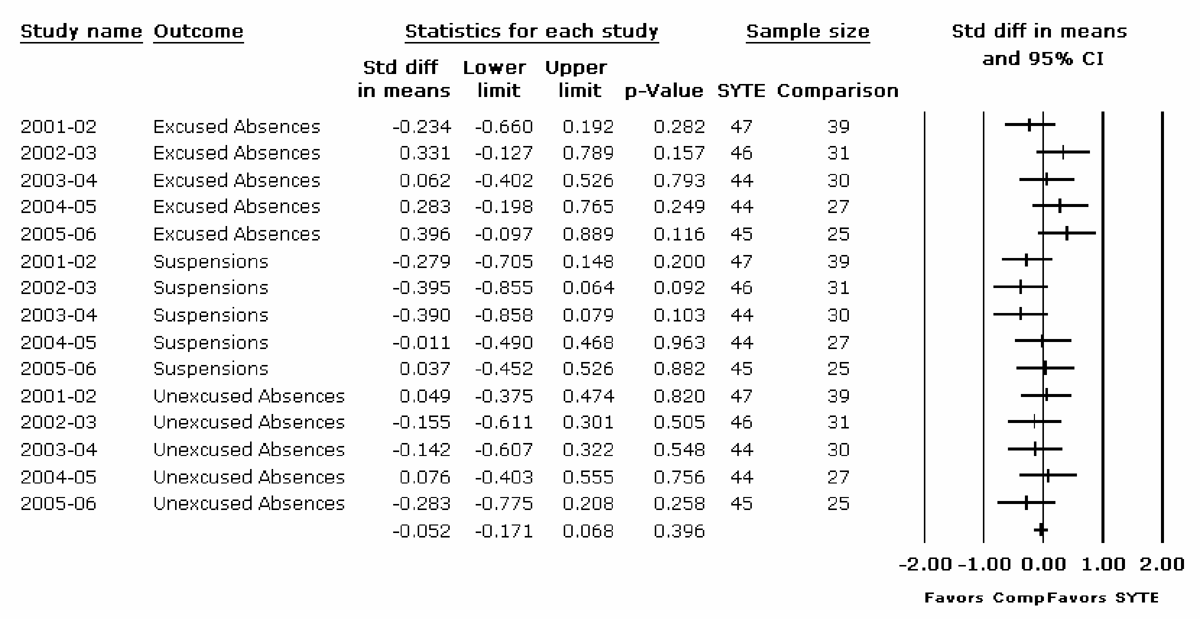
	Control					Say Yes				
	Promoted		Retained		Total	Promoted		Retained		Total
School Year	n	%	n	%	n	n	%	n	%	n
2000-01	45	100.0	0	0.0	45	45	100.0	0	0.0	45
2001-02	39	100.0	0	0.0	39	47	100.0	0	0.0	47
2002-03	30	96.8	1	3.2	31	44	95.7	2	4.3	46
2003-04	26	86.7	4	13.3	30	39	88.6	5	11.4	44
2004-05	21	77.8	6	22.2	27	44	100.0	0	0.0	44
2005-06	22	88.0	3	12.0	25	44	97.8	1	2.2	45
Total	183	92.9	14	7.1	197	263	97.0	8	3.0	271

Table 6 shows that promotion rates were slightly higher for SYTE participants relative to the comparison group (97% promoted on average compared to 93% promoted on average for the comparison group). Table 6 also shows that after 2002-03, the grade retention rates for the comparison group were 10% retained or higher where with the exception of one year (2003-04) SYTE retention rates were never higher than 4% of students retained. An important caveat to this comparison is that the proportion of retained students in the comparison group is also affected by the decreasing number of students in the cohort as shown in the Total column of Table 6. By 2005-06, there were only 25 students remaining in the comparison cohort whereas there SYTE remained relatively intact with 45 students.

4.5 Behavioral Outcomes: Attendance & Suspensions

SYTE students did not have significantly fewer absences overall than the comparison group. In fifth grade, for example, SYTE students averaged 12 absences while the comparison group averaged fifteen. However, Figure 5 shows, in 2005-06 the **SYTE group had an *educationally meaningful* lower average "Unexcused Absences"** than the comparison groups as measured in standard deviations (SDs) by the effect size (d = -0.28 SDs, 95%CI = -0.78 to 0.21) .

Figure 5. Comparison of SYTE and Comparison Groups on Behavioral Outcomes: 2000 - 02 through 2005 - 06.



Study name	Outcome	Statistics for each study				Sample size	
		Std diff in means	Lower limit	Upper limit	p-Value	SYTE	Comparison
2001-02	Excused Absences	-0.234	-0.660	0.192	0.282	47	39
2002-03	Excused Absences	0.331	-0.127	0.789	0.157	46	31
2003-04	Excused Absences	0.062	-0.402	0.526	0.793	44	30
2004-05	Excused Absences	0.283	-0.198	0.765	0.249	44	27
2005-06	Excused Absences	0.396	-0.097	0.889	0.116	45	25
2001-02	Suspensions	-0.279	-0.705	0.148	0.200	47	39
2002-03	Suspensions	-0.395	-0.855	0.064	0.092	46	31
2003-04	Suspensions	-0.390	-0.858	0.079	0.103	44	30
2004-05	Suspensions	-0.011	-0.490	0.468	0.963	44	27
2005-06	Suspensions	0.037	-0.452	0.526	0.882	45	25
2001-02	Unexcused Absences	0.049	-0.375	0.474	0.820	47	39
2002-03	Unexcused Absences	-0.155	-0.611	0.301	0.505	46	31
2003-04	Unexcused Absences	-0.142	-0.607	0.322	0.548	44	30
2004-05	Unexcused Absences	0.076	-0.403	0.555	0.756	44	27
2005-06	Unexcused Absences	-0.283	-0.775	0.208	0.258	45	25
		-0.052	-0.171	0.068	0.396		

Figure 5 also shows that except for two years (2001-02 and 2003-04) the SYTE group exhibited *educationally meaningful* higher mean numbers of excused absences and these results were sustained up to and including 2005-06. Providing excuses for absences are suggestive of responsibility and valid reasons for missing school. This is a positive outcome in contrast to having unexcused absences. In all years except the last two (2004-05 and 2005-06), the SYTE group also had *educationally meaningful* lower average number of suspensions than the comparison group (e.g., in 2003-04: d = -0.40, 95%CI = -0.86 to -.08). The mean number of suspensions, excused absences, unexcused absences, and other absences for the SYTE and comparison groups, for each school year, are presented in Appendix E.

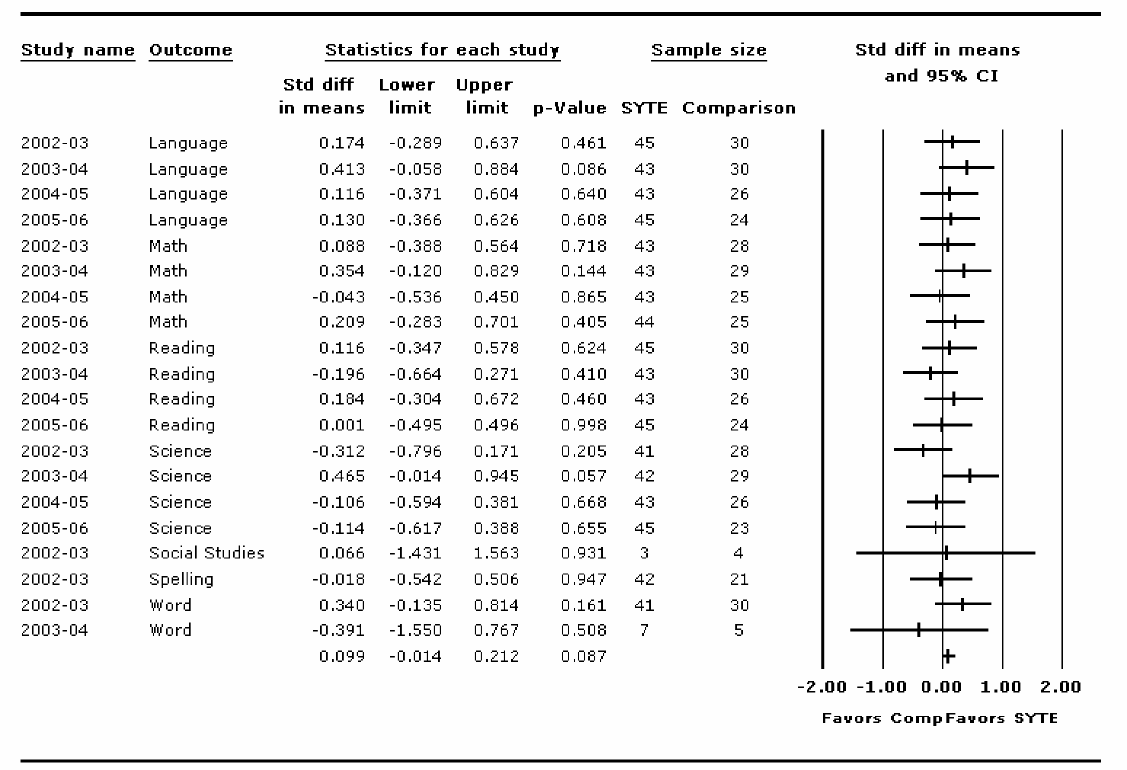
Follow-up analyses were conducted to determine whether there were any *educationally meaningfully* interaction effects between participation in the SYTE and student gender. For the 2005-06 suspension and excused absences data, SYTE males, on average, had a higher mean number of suspensions than comparison males and SYTE females, on average, had a higher mean number of excused absences than comparison females. However, the effect sizes were not educationally meaningful (i.e., were much smaller than the threshold of d = 0.25 SDs).

4.6 Academic Outcomes: Terra Nova Scale Scores

Figure 6 revealed an interesting trend in the performance of the SYTE group relative to the comparison group on Terra Nova scale scores. (Note: Terra Nova tests begin in second grade. **The SYTE outperformed the comparison group in core academic subjects in 2003-04 (when most students were in third grade), but in no other year.**

- Language Arts: d = 0.41 SDs, 95%CI = -0.06 to 0.88
- Math: d = 0.35 SDs, 95%CI = -0.83 to 0.12
- Reading: d = 0.24 SDs, 95%CI = -0.23 to 0.71
- Science: d = 0.46 SDs, 95%CI = -0.02 to 0.94

Figure 6. Comparison of SYTE and Comparison Groups on Terra Nova Scale Scores: 2002 - 03 through 2005 - 06.[16]



Study name	Outcome	Statistics for each study				Sample size		Std diff in means and 95% CI
		Std diff in means	Lower limit	Upper limit	p-Value	SYTE	Comparison	
2002-03	Language	0.174	-0.289	0.637	0.461	45	30	
2003-04	Language	0.413	-0.058	0.884	0.086	43	30	
2004-05	Language	0.116	-0.371	0.604	0.640	43	26	
2005-06	Language	0.130	-0.366	0.626	0.608	45	24	
2002-03	Math	0.088	-0.388	0.564	0.718	43	28	
2003-04	Math	0.354	-0.120	0.829	0.144	43	29	
2004-05	Math	-0.043	-0.536	0.450	0.865	43	25	
2005-06	Math	0.209	-0.283	0.701	0.405	44	25	
2002-03	Reading	0.116	-0.347	0.578	0.624	45	30	
2003-04	Reading	-0.196	-0.664	0.271	0.410	43	30	
2004-05	Reading	0.184	-0.304	0.672	0.460	43	26	
2005-06	Reading	0.001	-0.495	0.496	0.998	45	24	
2002-03	Science	-0.312	-0.796	0.171	0.205	41	28	
2003-04	Science	0.465	-0.014	0.945	0.057	42	29	
2004-05	Science	-0.106	-0.594	0.381	0.668	43	26	
2005-06	Science	-0.114	-0.617	0.388	0.655	45	23	
2002-03	Social Studies	0.066	-1.431	1.563	0.931	3	4	
2002-03	Spelling	-0.018	-0.542	0.506	0.947	42	21	
2002-03	Word	0.340	-0.135	0.814	0.161	41	30	
2003-04	Word	-0.391	-1.550	0.767	0.508	7	5	
		0.099	-0.014	0.212	0.087			

As Figure 6 also shows, the *educationally meaningful* effects exhibited by the SYTE group in 2003-04 were not sustained in subsequent years.[17] It is important to note that both the fourth and fifth grade Terra Nova exams were testing fourth grade learning. The SDP changed the timing of the Terra Nova exams to make the Terra Nova a diagnostic exam. Therefore, the fourth grade Terra Nova exam was given in the spring of the fourth grade year, the fifth grade Terra Nova was given in the fall of the fifth grade year. The fourth grade year was a particularly difficult one for SYTE students because they experienced the closing of the SYTE resource room and a long-term

[16] Results for Social Studies, Spelling, and Word are not reported for all years because either the sample sizes were too small, or the missing data was so substantial, that the scores were not available.

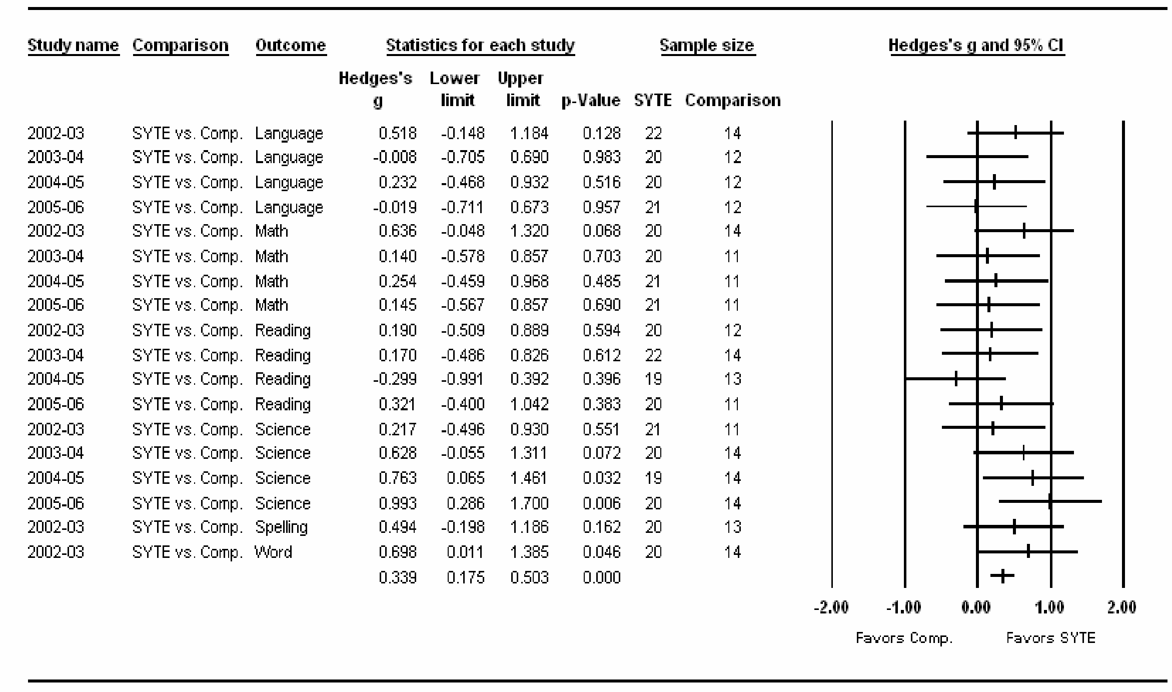
[17] The attrition of the comparison group could have biased the results in favor of SYTE. To address this issue, we tested the sensitivity of the effect sizes to comparison group attrition by computing an effect size for only those SYTE students with a matched comparison student that was still in the district administrative data file and had a valid score on the Terra Nova math test. The general trend in effect sizes remained the same. That is, for 2003-2004 and 2005-2006 specifically, the magnitude of the effect sized decreased slightly but in the former was still educationally meaningful and for the later it was still positive. Given these results on the math Terra Nova, we did not conduct a sensitivity analysis on other outcome variables. The results of this analysis are displayed in Appendix H.

substitute teacher for one fourth grade class. The means, standard deviations, and sample sizes for the SYTE and comparison groups are reported in Appendix F.

Mean comparisons, using effect sizes, were also conducted to test whether participation in the Say Yes program had differing effects on performance of males and females on the Terra Nova. Figure 7a displays this comparison for females in the SYTE program and females in the comparison group while Figure 7b displays the same comparison but for males. **Figure 7a shows that SYTE females outperformed comparison females in math each year** but by an educationally meaningful difference for 2002-03 and 2004-05 with d = 0.63 SDs, 95%CI = -0.05 to 1..32 and d = 0.25 SDs, 95%CI=-0.46 to 0.97, respectively. **Interestingly, comparison females outperformed SYTE females by an educationally meaningful difference in reading in 2004-05 (d = -0.30 SDs, 95%CI = -0.99 to 0.39) but the following year the SYTE females outperformed the comparison females in reading by an educationally meaningful difference (d = 0.32 SDs, 95%CI = -0.40 to 1.04).** Relative to girls in the comparison group, the performance of SYTE females in science was especially strong to the point that from 2003-04 on, the d indices were both educationally meaningful and in 2004-05 and 2005-06 the indices were also statistically significant. **In other words, the SYTE girls outperformed girls in the comparison group in science to the point that results were deemed as not due to chance—even with the same sample sizes that constrained statistical power.**[18]

[18] Although statistical reasoning leads us to conclude that the results were not due to chance, the differential attrition between the SYTE and comparison groups of girls are a threat to the internal validity of the study and cause us to wonder how much of the effect is due to the SYTE and how much is due to the less academically able girls in the comparison group leaving the School District or not taking the Terra Nova.

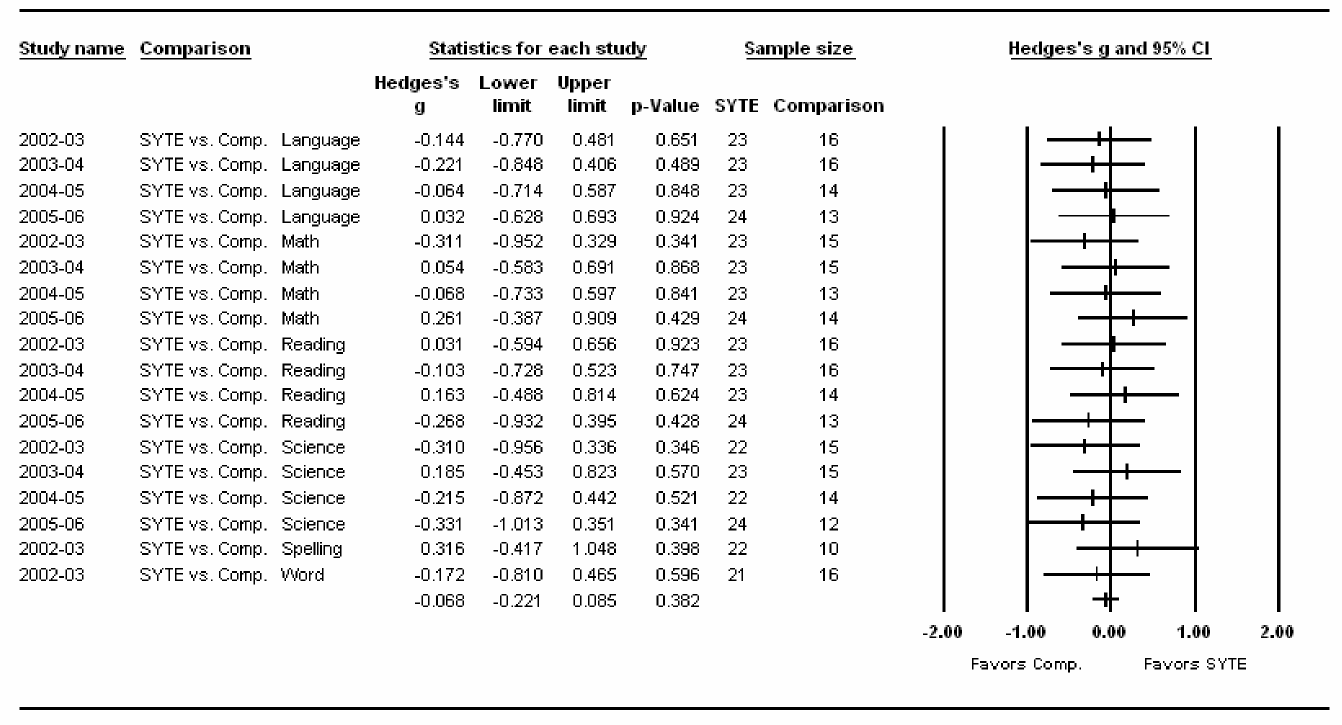
Figure 7a. Comparison of SYTE Females and Comparison Females by on Terra Nova Scale Scores: 2000 - 01 through 2005 - 06.



Study name	Comparison	Outcome	Statistics for each study				Sample size	
			Hedges's g	Lower limit	Upper limit	p-Value	SYTE	Comparison
2002-03	SYTE vs. Comp.	Language	0.518	-0.148	1.184	0.128	22	14
2003-04	SYTE vs. Comp.	Language	-0.008	-0.705	0.690	0.983	20	12
2004-05	SYTE vs. Comp.	Language	0.232	-0.468	0.932	0.516	20	12
2005-06	SYTE vs. Comp.	Language	-0.019	-0.711	0.673	0.957	21	12
2002-03	SYTE vs. Comp.	Math	0.636	-0.048	1.320	0.068	20	14
2003-04	SYTE vs. Comp.	Math	0.140	-0.578	0.857	0.703	20	11
2004-05	SYTE vs. Comp.	Math	0.254	-0.459	0.968	0.485	21	11
2005-06	SYTE vs. Comp.	Math	0.145	-0.567	0.857	0.690	21	11
2002-03	SYTE vs. Comp.	Reading	0.190	-0.509	0.889	0.594	20	12
2003-04	SYTE vs. Comp.	Reading	0.170	-0.486	0.826	0.612	22	14
2004-05	SYTE vs. Comp.	Reading	-0.299	-0.991	0.392	0.396	19	13
2005-06	SYTE vs. Comp.	Reading	0.321	-0.400	1.042	0.383	20	11
2002-03	SYTE vs. Comp.	Science	0.217	-0.496	0.930	0.551	21	11
2003-04	SYTE vs. Comp.	Science	0.628	-0.055	1.311	0.072	20	14
2004-05	SYTE vs. Comp.	Science	0.763	0.065	1.461	0.032	19	14
2005-06	SYTE vs. Comp.	Science	0.993	0.286	1.700	0.006	20	14
2002-03	SYTE vs. Comp.	Spelling	0.494	-0.198	1.186	0.162	20	13
2002-03	SYTE vs. Comp.	Word	0.698	0.011	1.385	0.046	20	14
			0.339	0.175	0.503	0.000		

In contrast, Figure 7b shows that SYTE males rarely outperformed boys in the comparison group by an educationally meaningfully difference with a few important exceptions: in 2002-03 in spelling and 2005-06 in math where d = 0.32 SDs, 95%CI=-0.42 to 1.05 and d = 0.26 SDs, 95%CI = -0.39 to 0.91, respectively. Figure 7b also shows that in 2005-06, males in the comparison group outperformed males in the SYTE program by educationally meaningful differences in reading and science. Taken together, the results for SYTE males and SYTE females suggest that there are gender interactions for the SYTE program. Stated differently, the SYTE program may have differential effects on females and males. The means, standard deviations, and samples for gender subgroup comparisons on the Terra Nova are presented in Appendix H.

Figure 7b. Comparison of SYTE Males and Comparison Males on Terra Nova Scale Scores: 2000 - 01 through 2005 - 06.



Study name	Comparison		Statistics for each study Hedges's g	Lower limit	Upper limit	p-Value	Sample size SYTE	Comparison
2002-03	SYTE vs. Comp.	Language	-0.144	-0.770	0.481	0.651	23	16
2003-04	SYTE vs. Comp.	Language	-0.221	-0.848	0.406	0.489	23	16
2004-05	SYTE vs. Comp.	Language	-0.064	-0.714	0.587	0.848	23	14
2005-06	SYTE vs. Comp.	Language	0.032	-0.628	0.693	0.924	24	13
2002-03	SYTE vs. Comp.	Math	-0.311	-0.952	0.329	0.341	23	15
2003-04	SYTE vs. Comp.	Math	0.054	-0.583	0.691	0.868	23	15
2004-05	SYTE vs. Comp.	Math	-0.068	-0.733	0.597	0.841	23	13
2005-06	SYTE vs. Comp.	Math	0.261	-0.387	0.909	0.429	24	14
2002-03	SYTE vs. Comp.	Reading	0.031	-0.594	0.656	0.923	23	16
2003-04	SYTE vs. Comp.	Reading	-0.103	-0.728	0.523	0.747	23	16
2004-05	SYTE vs. Comp.	Reading	0.163	-0.488	0.814	0.624	23	14
2005-06	SYTE vs. Comp.	Reading	-0.268	-0.932	0.395	0.428	24	13
2002-03	SYTE vs. Comp.	Science	-0.310	-0.956	0.336	0.346	22	15
2003-04	SYTE vs. Comp.	Science	0.185	-0.453	0.823	0.570	23	15
2004-05	SYTE vs. Comp.	Science	-0.215	-0.872	0.442	0.521	22	14
2005-06	SYTE vs. Comp.	Science	-0.331	-1.013	0.351	0.341	24	12
2002-03	SYTE vs. Comp.	Spelling	0.316	-0.417	1.048	0.398	22	10
2002-03	SYTE vs. Comp.	Word	-0.172	-0.810	0.465	0.596	21	16
			-0.068	-0.221	0.085	0.382		

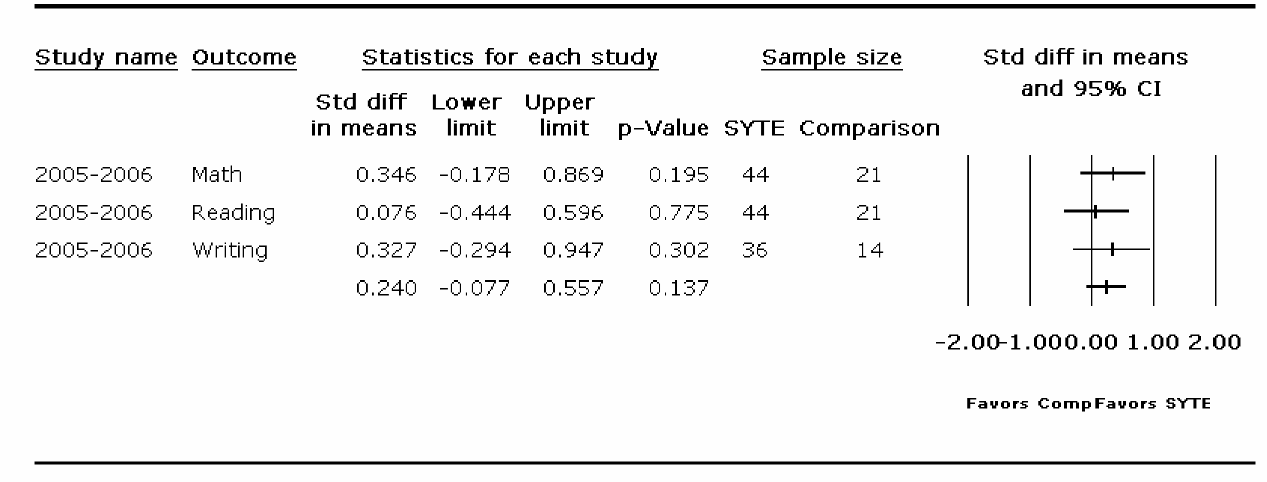
4.7 Academic Outcomes: PSSA Scales Scores and Criterion

The Pennsylvania System of School Assessment (PSSA) is a standards-based criterion-referenced assessment used to measure student attainment of academic standards while also determining the degree to which school programs enable students to attain proficiency in meeting those standards. An important difference between the PSSA and Terra Nova is that the former is administered in the spring near of the end of the school year and the latter is administered in the fall at the beginning of the school year. Beginning in the 2006-07 school year, every Pennsylvania student in grades 3 through 8 and grade 11 is assessed in reading and math, while students in grades 5, 8 and 11 are assessed in writing (PA DOE website, January 15, 2007). However, in this analysis we report the group comparison on only 2005-06 PSSA scores because these were the only scores available in the administrative data file provided by the SDP.

For the Spring 2006 administration of the PSSA, SYTE and comparison students on a normal academic progression would have been in the 5th grade. As previously shown in Tables 5 and 6, not all Say Yes or comparison students were in 5th grade in 2005-06. Thus, analysis sample sizes for comparing scale scores on the PSSA Reading, Mathematics, and Writing tests will be slightly less than the baseline samples or even sample size in previous years.

Figure 8 shows that SYTE group outperformed the comparison group in writing and math as measured by *educationally meaningful* effect sizes of d = 0.35 (95%CI = -0.18 to 0.87) and d = 0.33 (95%CI=-0.29 to 0.95), respectively. The means, standard deviations, and sample sizes for the group comparisons are presented in Appendix I. Tests for educationally meaningful differences between SYTE and the comparison group in reading did not show differences. In addition, there were no educationally meaningful differences between females in the SYTE and comparison groups and males in the SYTE and comparison groups on any of the PSSA tests.

Figure 8. Comparison of SYTE and Comparison Groups on PSSA Scale Scores for 2005 06.



Study name	Outcome	Statistics for each study				Sample size	
		Std diff in means	Lower limit	Upper limit	p-Value	SYTE	Comparison
2005-2006	Math	0.346	-0.178	0.869	0.195	44	21
2005-2006	Reading	0.076	-0.444	0.596	0.775	44	21
2005-2006	Writing	0.327	-0.294	0.947	0.302	36	14
		0.240	-0.077	0.557	0.137		

An alternative way to examine PSSA scores is by level of achievement for each subject. An achievement level of three (3) or greater is considered proficient by PA DOE standards. **For Math, Reading, and Writing, SYTE students were more likely to attain a level 3, reaching proficiency as defined by the PSSA cut-scores.** Complete results are presented in Appendix J.

4.8 Academic Outcomes: Grades

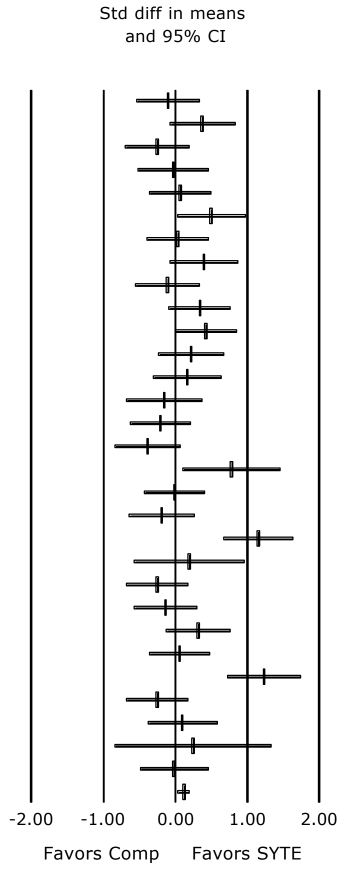
In first grade, SYTE students were rated higher in “knowledge of number systems” (d=0.42 standard deviations) and “nature of science” (d=0.34 standard deviations). However, they were rated lower (grades differences were educationally meaningful) in “social studies skills” (d= -0.25 standard deviations) and “work habits” (d=0.25 standard deviations).

In second grade, SYTE students were rated higher in two out of four mathematical skill areas (results, educationally meaningful, d =0.38 standard deviations, d = 0.26 standard deviations). They were also rated much higher in both instructional reading and independent reading (d= 0.40 standard deviations, d= 0.51 standard deviations)**. Finally, their grade average was one full standard deviation higher in both “stages of writing” and science which is equivalent to a full school year of academic growth.**

No grade differences were observed between SYTE and the comparison group in third and fourth grades. An analysis of grade differences in fifth grade was not conducted because many of the students attend KIPP Academy which uses a different report card format making it difficult to make comparisons across schools. Grades are, of course, teacher assessments and more subjective than standardized tests. However, the positive results in the early grades is consistent with the results on the Terra Nova. Complete results for student grades as outcomes are presented in Appendix K.

Figure 9. Comparison of SYTE and Comparison Groups on Grades for Available Years

Study name	Outcome	Statistics for each study				Sample size	
		Std diff in means	Lower limit	Upper limit	p-Value	SYTE	Comparison
2001-02	Data, Stats, & Probability	-0.101	-0.547	0.344	0.656	47	33
2002-03	Data, Stats, & Probability	0.377	-0.084	0.839	0.109	45	31
2002-03	Geometry	-0.254	-0.713	0.206	0.279	45	31
2003-04	Independent Reading	-0.027	-0.518	0.464	0.913	44	25
2001-02	Independent Reading	0.075	-0.360	0.509	0.737	47	36
2002-03	Independent Reading	0.502	0.019	0.985	0.042	43	28
2001-02	Instructional Reading	0.030	-0.404	0.465	0.891	47	36
2002-03	Instructional Reading	0.402	-0.078	0.883	0.101	43	28
2002-03	Measurement	-0.111	-0.569	0.347	0.635	45	31
2001-02	Nature of Science	0.342	-0.091	0.776	0.122	47	37
2001-02	Number Systems	0.424	-0.012	0.859	0.056	47	37
2002-03	Number Systems	0.221	-0.238	0.680	0.346	45	31
2003-04	Other Curricular Area	0.171	-0.315	0.657	0.490	44	26
2004-05	Other Curricular Area	-0.151	-0.684	0.382	0.579	42	20
2001-02	Other Curricular Area	-0.204	-0.636	0.228	0.355	47	37
2002-03	Other Curricular Area	-0.381	-0.850	0.087	0.110	44	30
2005-06	Other Curricular Area Course	0.783	0.103	1.463	0.024	22	15
2001-02	Patterns, Algebra & Functions	-0.008	-0.443	0.426	0.970	47	36
2002-03	Patterns, Algebra & Functions	-0.190	-0.649	0.268	0.416	45	31
2002-03	Science Courses	1.156	0.663	1.649	0.000	45	31
2003-04	Service Learning Project	0.194	-0.575	0.963	0.621	35	8
2001-02	Social Skills Marks	-0.248	-0.681	0.184	0.260	47	37
2001-02	Social Studies	-0.136	-0.579	0.308	0.549	46	34
2002-03	Social Studies	0.317	-0.143	0.777	0.177	45	31
2001-02	Stages of Writing	0.062	-0.369	0.493	0.779	47	37
2002-03	Stages of Writing	1.235	0.717	1.752	0.000	43	28
2001-02	Work Habits Marks	-0.249	-0.681	0.184	0.260	47	37
2003-04	Writing Comp	0.099	-0.392	0.590	0.692	44	25
2005-06	Writing Comp	0.252	-0.842	1.347	0.651	6	7
2002-03	Writing Composition	-0.013	-0.489	0.463	0.958	43	28
		0.119	0.032	0.206	0.007		



4.9 Multi-variate Longitudinal Investigation of Academic Outcomes

To extend the longitudinal analysis beyond examining the yearly trends in effect sizes for behavioral and academic outcomes previously presented in the Forrest Plots, we endeavored to model the amount of yearly change in effect sizes through multi-level growth models in which academic outcomes are nested within students. As recommended by Singer & Willet (2003), as an initial step we estimated an unconditional growth model, where time is the only predictor associated with the criterion variable which in this case was the Terra Nova and PSSA scores. The unconditional model provides the benchmark by which subsequent models, including other predictors of interest, are evaluated. Chief among these predictors is the inspection of the parameter representing time. There was insufficient variance in the criterion variable across time to proceed with the analysis. This was the case for both the behavioral and academic outcome variables.

V. Conclusions

We return now to address the research questions posed earlier. These questions were as follows:

1. Are SYTE students on track to graduate from high school?

2. What is the discernible impact of the Say Yes program on participants' scores on standardized achievement tests, promotion rates, grades, attendance and behavior marks annually and if there is a discernible impact, does it vary by gender?

3. If a discernible impact of the Say Yes program exists, does the impact vary over time (from the first year of the program to the fifth)?

While behavioral and academic performance in fifth grade has not been found to be a reliable predictor of whether students will graduate or drop out, SYTE students appear to be on-track to graduate according to most indicators. Eighty-five percent of SYTE students are in the appropriate grade and have not been retained. They attend school regularly and do not exhibit significant behavioral problems. However, only 25-30% is performing on grade level in math and reading, according to the state standardized tests. This is cause for concern as they move forward.

At the same time, SYTE students performed better than a matched comparison group in a number of areas and in some years. They outperformed the comparison group most consistently and to the greatest degree on the third grade Terra Nova language arts, reading, science and math exams. They also outperformed the comparison group on the math and writing state standardized PSSA tests in the spring of their fifth grade year, the first year in a new middle school for many of the students. The group performance was bolstered by students who attended KIPP Academy charter school. This finding is important because students often lose academic ground after a transition. However, SYTE students appear to have successfully weathered this transition. This finding also raises an interesting question, beyond the scope of this study, of the contribution of the school compared to SYTE ancillary academic supports to SYTE students' academic performance.

Females and males were impacted differently by SYTE. The impact was strongest for girls and manifested itself in math and reading Terra Nova exams and was most pronounced and sustained in the Terra Nova science exam. The effect of the SYTE program on girls was so strong in science in 2003-04 and beyond that not only were the effect sizes educationally meaningful but they were also statistically significant even with the small sample sizes. The effect of the SYTE program worked in the opposite direction for males except in spelling in second grade and encouragingly, in math in fifth grade. With the national debate on appropriate policies to address the increasing marginalization of African American males in the United States, finding ways to leverage the impact the SYTE program such that it extends to males is an important issue to investigate programmatically.

SYTE students also seemed to have experienced a more stable and supportive context for learning than the comparison group. SYTE students were more likely than the comparison group to remain in the school district suggesting they had more continuity and stability in their educational experience. They were more likely to receive supports for special education and giftedness than the comparison group, and their parents were more likely to provide the school with reasons for their absences suggesting that the parents were more engaged with the school than parents of comparison group children.

Student behavioral outcomes showed some small effects from SYTE in elementary school. SYTE total number of absences did not differ from the comparison group for any year of the analysis but they did have fewer suspensions than the comparison group in second and third grades.

VI. Limitations of this study and recommendations for future research

No study has perfect validity and this study is no exception. There are several limitations of this study which will be discussed below. Some of these limitations are artifacts of the program stage and model and were addressed in the research design. Other limitations were unanticipated and could be corrected in future research.

First, a randomized controlled trial was impossible to implement in this study given that students were selected for the program in kindergarten and this evaluation design was developed when they were entering 6th grade. Without random assignment, there are always reservations about how much the effect size can be attributed to the SYTE program because unlike a high-quality randomized controlled trial, this quasi-experimental design did not necessarily balance the groups being compared on unobservable characteristics such as attitude, motivation, achievement orientation and the like.

Second, the study was woefully underpowered from the outset ($b \cong 0.17$) and the power could not be increased because the evaluation was conducted post-hoc. Increasing the SYTE Group to comparison group matching ratio to 1:8 (or 45 SYTE participants to 360 comparison students) increase powered only up to 0.21 but lead to difficulties in findings matches in the comparison pool even though the pool comprised more than 23,000 kindergartners (recall the matching was done on 17 covariates which made findings a matched based on the propensity score challenging even with the large sample). For these reasons, we stayed with the 1:1 propensity score matching ratio resulting in balanced groups of 45 kindergartners in 2000-01, reported effect sizes with confidence intervals and p-values, and acknowledged the relevance of educationally meaningful effect size separate and distinct from statistical significance which depends heavily on sample size.

One reason for the difficulty finding appropriate matches from such a large pool was that one of the covariates, zip code, which served as a proxy for neighborhood effects, reduced the pool to just a few thousand. Therefore, the use of zip code resulted in an extremely conservative matching. Future research should find other ways of controlling for neighborhood effects without using zip code.

There was differential attrition in this study spurred by the lack of stability in the comparison group and the stability of the SYTE group. The stability in the SYTE group was expected but the lack of stability in the comparison group was not. By 2005-06, the comparison group comprised 25 out of the original 45 students (or 44% attrition). Programmatically, this result speaks to the SYTE program's positive "cohort effect" that maintains the cohesiveness of a group such that they remain in the district and proceed through the education pipeline, for the most part, on time. Methodologically, the 44% attrition in the comparison leaves us to wonder if the results would still hold if we could conduct an "intent to treat" analysis in which those outcomes for the comparison students that were no longer in the administrative data file could be included in the analysis.

We tested the sensitivity of the effect sizes to comparison group attrition by computing an effect size for only those SYTE students with a matched comparison student that was still in the district administrative data file and had a valid score on the Terra Nova. The general trend in the effect sizes remained the same and therefore, we did not conduct a sensitivity analysis for other outcome variables. If future research is similarly affected by attrition, a sensitivity analysis should be conducted on all outcome variables. Future research should also request school entry, withdrawal codes and dates so that analysts can investigate why students, especially those in the comparison group, left a particular school or the SDP entirely. Ideally, future research would also provide the resources so that comparison group students leaving the district could continue to be tracked and compared to the SYTE group.

This study by design, and because of the limited types of measures available through administrative records, focus on a narrow set of outcomes, namely standardized academic achievement and a very limited number of behavioral outcomes in the form of absences and suspensions. It is conceivable that SYTE could impact a number of psycho-social outcomes such as self-esteem, locus of control, and attitudes towards learning—none of which were outcomes in this evaluation.

While the database that was developed for this study has limitations, it is extensive and SYTE should build upon it for future research. Given the challenges with missing data in the kindergarten grade files and the smaller pool created when zip code is included as a covariate for matching, a new comparison group could be created matching in first grade and dropping zip code from the matching process. School attended (focusing on the lowest performing eighty-six schools that were involved in the SDP reform efforts) may provide an adequate proxy for neighborhood and should expand the pool for selecting a comparison group. A larger comparison pool would then allow multiple matches for each SYTE student to be identified and this would help to address the challenge of attrition from the comparison group as well as increasing the statistical power of the analysis. Outcome data could be added to the existing file each year and SYTE student progress could easily be compared to a matched group.

An important programmatic question suggested by this analysis, which future research should explore is the relative contribution of the school to student outcomes. The strong performance of students who left the original neighborhood elementary school for charter and magnet schools suggests that ancillary academic supports are not enough, at least for males, to compensate for a weak school environment. A comparison between the fall 05-06 Terra Nova scores and the fall 06-07 Terra Nova scores for Philadelphia students could confirm whether KIPP academy has indeed boosted SYTE students' academic performance. Research in other SYTE programs, which have larger sample sizes, could also test the contribution of the school to SYTE student outcomes.

References

Agodino, R., & Dynarski, M. (2004). Are experiments the only option? A look at dropout prevention programs. *The Review of Economics and Statistics*, *86*(1), 180-194.

Allison, P. (2001). *Missing Data.* Sage University Papers Series on Quantitative Applications in the Social Sciences, 07-136. Thousand Oaks, CA: Sage.

Borenstein, M. (2005). Software for publication bias. In H. Rothstein, A. Sutton, & M. Borenstein (Eds.), *Publication bias in meta-analysis* (pp. 193-220). West Sussex, England: Wiley.

Cochran, W. (1965). The planning of observational studies of human populations. *Journal of the Royal Statistical Society, Series A, 128,* 234-255.

Cochran, W. (1968). The effectiveness of adjustment by sub classification in removing bias in observational studies. *Biometrics, 24,* 205-213.

Dehejia, R. H., & Wahba, S. (1999). Causal effects in nonexperimental studies: Reevaluating the evaluation of training programs. *Journal of the American Statistical Association*, *94,* 1053–1062.

Glazerman, S., Levy, D. M., & Myers, D. (2003). Nonexperimental versus experimental estimates of earnings impacts. *The ANNALS of the American Academy of Political and Social Science, 589*, 63–93.

Luellen, J., Shadish, W., & Clark, M. (2005). Propensity scores: An introduction and experimental test. *Evaluation Review, 29*(6) 530-558.

Pasta, D. (2000). *Using propensity scores to adjust for group differences: Examples comparing alternative surgical methods.* Paper presented at the 25th Annual SAS Users Group International Conference. Indianapolis, IN.

Rosenbaum, P. R. & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika, 70*(1) 41-55.

Rosenbaum, P. R. & Rubin, D. B. (1984). Reducing bias in observational studies using sub classification on the propensity score. *Journal of the American Statistical Association, 79,* 387, 516-524.

Singer, J. & Willet, J. (2003). *Applied Longitudinal Data Analysis: Modeling Change and Event Occurrence*. New York: Oxford University Press

Victor, T. W., & Boruch, R. F. (2007). *Examination of the Statistical Bias of Various Estimators.* Unpublished Doctoral Dissertation, University of Pennsylvania, Philadelphia. .

Appendix A: Study Variables and Availability from the School District of Philadelphia

Construct	Cataloged	Variable Name	School Years	"SYTE" Grade Cohorts
Baseline Covariates for Matched Pairs and Regression:				
Academic Achievement:				
Standard or Criterion Reference Tests:				
Pre-K test scores	No	-	-	-
Grades	Yes		2000-01	Kindergarten
Days Absence	Yes		2000-01	Kindergarten
Days Late	Yes		2000-01	Kindergarten
Students' family characteristics	No	-	-	-
Free and reduced lunch status	Yes		2000-01	Kindergarten
School readiness	No	-	-	-
Zip code (for matching in other data files)	Yes		2000-01	Kindergarten
Neighborhood characteristics	No	-	-	-
School type (i.e., 82 lowest performing vs. others)[19]	Yes		2000-01	Kindergarten
Ethnicity	Yes		2000-01	Kindergarten
Gender	Yes		2000-01	Kindergarten
English as a Second Language	Yes		2000-01	Kindergarten
Special Education	Yes		2001-06	1st -5th grade
Outcomes for Matched Pairs and Regression:				
Academic Achievement:				
Terra Nova: Reading, Language, Math, & Science	Yes		2003 - 06	2nd - 5th grade
Student Grades	Yes		2001-06	1st -5th grade
Student Attendance:				
Suspension, Excused Absences, Unexcused Absences, & Other Absences	Yes		2001-06	1st -5th grade
Teacher Ratings of Student Behavior:				
Social Skills	Yes		2002-03	2nd and 3rd grade
Work Habits	Yes		2002-03	2nd and 3rd grade
Student Retention in Grade	Yes		2001-06	1st -5th grade
Student School Transfer	No	-	-	-

[19] This number is lower than the original "86" because there were some kids in the data file did not have a school number of the school was closed.

Appendix B

Propensity match results for the Zip Code Variable.

	Comparison		SYTE		Total
Zip Code	n	%	n	%	n
19104		0.0	1	2.2	1
19121		0.0	1	2.2	1
19139	10	22.2	14	31.1	24
19143	35	77.8	28	62.2	63
19151		0.0	1	2.2	1
Total	45	100.0	45	100.0	90

Propensity match results for the School Variable.

	Comparison		SYTE		Total
School	n	%	n	%	n
Bryant	40	88.9	43	95.6	83
Harrity	2	4.4	1	2.2	3
Anderson	3	6.7	1	2.2	4
Total	45	100.0	45	100.0	90

Appendix C

Propensity Match Results for Attendance Variables

Variable	Statistics	Comparison	SYTE
Days Absent	N	45	45
	Missing	0	0
	Mean	49.0	49.8
	SD	5.4	4.2
	Minimum	34	40
	Maximum	55	55
Days Late	N	45	45
	Missing	0	0
	Mean	5.9	5.2
	SD	5.3	4.2
	Minimum	0	0
	Maximum	21	15

Appendix D

Propensity Match Results for Kindergarten Marks

Variable	Statistics	Comparison	SYTE
Math	N	45	45
	Missing	0	0
	Mean	9.3	7.6
	Minimum	1	1
	Maximum	31	33
	SD	7.6	7.2
Language Arts	N	45	45
	Missing	0	0
	Mean	10.8	9.6
	SD	8.1	6.6
	Minimum	2	2
	Maximum	30	42
Personal Growth	N	45	45
	Missing	0	0
	Mean	6.4	6.5
	SD	4.5	3.6
	Minimum	1	1
	Maximum	18	16
Work Habits	N	45	45
	Missing	0	0
	Mean	4.1	4.2
	SD	3.3	2.8
	Minimum	1	1
	Maximum	15	15

Appendix D (Continued)

Variable	Statistics	Comparison	SYTE
Physical Development	N	45	45
	Missing	0	0
	Mean	5.9	4.9
	SD	5.5	4.7
	Minimum	1	1
	Maximum	21	28
Art & Music	N	45	45
	Missing	0	0
	Mean	1.5	1.6
	SD	0.5	0.4
	Minimum	1	1
	Maximum	3	3
Science	N	45	45
	Missing	0	0
	Mean	2.4	2.6
	SD	2.1	1.4
	Minimum	1	1
	Maximum	9	9
Social Studies	N	45	45
	Missing	0	0
	Mean	4.4	4.1
	SD	2.8	3.5
	Minimum	1	1
	Maximum	12	15

Appendix E: Comparison of SYTE and Comparison Group on Behavioral Outcomes

		SYTE			Comparison			Total		
Year	Outcome	Mean	N	SD	Mean	N	SD	Mean	N	SD
2000-01	Absences	49.82	45	4.21	49.00	45	5.42	49.41	90	4.85
	Days Late	5.16	45	4.18	5.92	45	5.28	5.54	90	4.75
2001-02	Suspensions	0.00	47	0.00	0.03	39	0.16	0.01	86	0.11
	Excused Absences	3.26	47	3.53	6.95	39	23.13	4.93	86	15.79
	Unexcused Absences	10.13	47	9.43	10.59	39	9.22	10.34	86	9.28
	Other Absences	0.00	47	0.00	0.03	39	0.16	0.01	86	0.11
2002-03	Suspensions	0.00	46	0.00	0.16	31	0.64	0.06	77	0.41
	Excused Absences	13.26	46	28.73	5.68	31	8.55	10.21	77	23.06
	Unexcused Absences	18.13	46	15.67	20.94	31	21.26	19.26	77	18.05
	Other Absences	0.00	46	0.00	0.26	31	1.12	0.10	77	0.72
2003-04	Suspensions	0.05	44	0.21	0.20	30	0.55	0.11	74	0.39
	Excused Absences	4.45	44	5.96	4.13	30	3.66	4.32	74	5.12
	Unexcused Absences	6.57	44	6.02	7.57	30	8.30	6.97	74	7.00
	Other Absences	0.05	44	0.21	0.43	30	1.41	0.20	74	0.92
2004-05	Suspensions	0.18	44	1.06	0.19	27	0.48	0.18	71	0.88
	Excused Absences	6.16	44	6.92	4.30	27	5.94	5.45	71	6.58
	Unexcused Absences	8.86	44	9.41	8.15	27	9.28	8.59	71	9.30
	Other Absences	0.50	44	2.89	0.44	27	1.45	0.48	71	2.43
2005-06	Suspensions	0.27	45	0.86	0.24	25	0.72	0.26	70	0.81
	Excused Absences	4.38	45	4.79	2.56	25	4.22	3.73	70	4.65
	Unexcused Absences	8.00	45	10.88	12.24	25	20.43	9.51	70	15.00
	Other Absences	0.00	45	0.00	0.00	25	0.00	0.00	70	0.00

Appendix F: Comparison of SYTE and Comparison Group on Terra Nova Scores (Fall 2002 – Fall 2005)

		Say Yes			Comparison			Total		
Year	Outcome	Mean	N	SD	Mean	N	SD	Mean	N	SD
2002-03	Language	41.11	45	17.83	38.20	30	14.84	39.95	75	16.65
	Math	40.95	43	25.34	39.00	28	16.12	40.18	71	22.06
	Reading	41.96	45	16.92	40.13	30	14.02	41.23	75	15.75
	Science	27.29	41	14.95	32.07	28	15.82	29.23	69	15.37
	Social Studies	11.00	3	17.32	10.25	4	3.86	10.57	7	10.37
	Spelling	44.38	42	21.27	44.76	21	21.84	44.51	63	21.29
	Word	37.98	41	13.33	33.63	30	12.06	36.14	71	12.90
2003-04	Language	43.40	43	13.17	37.90	30	13.54	41.14	73	13.50
	Math	43.86	43	19.54	37.41	29	16.03	41.26	72	18.37
	Reading	41.95	43	15.01	39.10	30	13.74	40.78	73	14.47
	Science	38.93	42	12.76	33.03	29	12.57	36.52	71	12.93
	Social Studies				51.00	1	.	51.00	1	.
	Spelling				41.00	2	4.24	41.00	2	4.24
	Word	39.29	7	12.93	44.60	5	14.47	41.50	12	13.22
2004-05	Language	41.28	43	15.03	39.42	26	17.46	40.58	69	15.89
	Math	38.19	43	19.05	38.96	25	15.81	38.47	68	17.81
	Reading	38.60	43	15.40	36.00	26	11.74	37.62	69	14.10
	Science	33.40	43	13.14	34.73	26	11.31	33.90	69	12.41
	Social Studies	1.00	1	.	18.43	7.00	21.00	16.25	8	20.39
	Spelling	-	-	-	-	-	-	-	-	-
	Word				31.00	1	-	31.00	1	-
2005-06	Language	41.87	45	19.38	39.54	24	14.84	41.06	69	17.85
	Math	46.23	44	20.28	42.16	25	17.90	44.75	69	19.42
	Reading	40.47	45	16.93	40.46	24	15.24	40.46	69	16.25
	Science	29.84	45	15.41	31.43	23	10.19	30.38	68	13.81
	Social Studies	1.00	1	-	27.00	2	12.73	18.33	3	17.50
	Spelling	-	-	-	-	-	-	-	-	-
	Word	-	-	-	-	-	-	-	-	-

Appendix G: Sensitivity Analysis with Terra Nova Math Score

TN Math Revisit

		Say Yes			Comparison						
Year	Outcome	Mean	N	SD	Mean	N	SD	d	SE	Lower CI	Upper CI
2002-03	Math	37.36	25	27.26	38.68	25	16.69	-0.058	0.283	-0.613	0.496
2003-04	Math	43.21	24	20.16	37.63	24	17.45	0.296	0.290	-0.273	0.865
2004-05	Math	37.86	21	20.80	38.33	21	17.18	-0.025	0.309	-0.630	0.580
2005-06	Math	45.35	23	20.47	42.04	23	18.62	0.169	0.295	-0.410	0.748

Appendix H: Gender Interactions on Terra Nova Scores

		SYTE - Females				Comparison - Females		
Year	Domain	Mean	SD	n		Mean	SD	n
2002-03	Language	49.27	16.75	22		41.14	12.80	14
2003-04	Language	47.05	12.91	20		38.79	12.33	14
2004-05	Language	48.05	13.65	20		45.33	14.43	12
2005-06	Language	46.90	20.37	21		42.64	16.40	11
2002-03	Math	49.05	26.87	20		37.38	15.14	13
2003-04	Math	46.40	20.26	20		34.00	11.84	14
2004-05	Math	41.35	21.87	20		41.50	12.26	12
2005-06	Math	50.50	25.02	20		47.45	11.08	11
2002-03	Reading	48.73	16.95	22		45.93	14.61	14
2003-04	Reading	47.85	14.45	20		39.86	8.64	14
2004-05	Reading	45.20	13.03	20		42.25	11.22	12
2005-06	Reading	45.52	18.72	21		40.91	15.33	11
2002-03	Science	27.79	17.60	19		33.31	18.52	13
2003-04	Science	41.16	12.97	19		31.29	12.12	14
2004-05	Science	35.81	15.01	21		36.08	11.02	12
2005-06	Science	32.76	15.00	21		30.73	10.50	11
2002-03	Spelling	52.65	19.92	20		45.82	22.17	11
2002-03	Word	44.75	10.98	20		33.50	11.18	14

		SYTE - Males				Comparison - Males		
Year	Domain	Mean	SD	n		Mean	SD	n
2002-03	Language	33.30	15.40	23		35.63	16.39	16
2003-04	Language	40.22	12.81	23		37.13	14.87	16
2004-05	Language	35.39	13.88	23		34.36	18.72	14
2005-06	Language	37.46	17.73	24		36.92	13.48	13
2002-03	Math	33.91	22.16	23		40.40	17.33	15
2003-04	Math	41.65	19.07	23		40.60	19.01	15
2004-05	Math	35.43	16.20	23		36.62	18.70	13
2005-06	Math	42.67	14.92	24		38.00	21.32	14
2002-03	Reading	35.48	14.45	23		35.06	11.66	16
2003-04	Reading	36.83	13.81	23		38.44	17.30	16
2004-05	Reading	32.87	15.24	23		30.64	9.56	14
2005-06	Reading	36.04	14.14	24		40.08	15.78	13
2002-03	Science	26.86	12.64	22		31.00	13.64	15
2003-04	Science	37.09	12.57	23		34.67	13.18	15
2004-05	Science	31.09	10.92	22		33.57	11.83	14
2005-06	Science	27.29	15.63	24		32.08	10.33	12
2002-03	Spelling	36.86	20.00	22		43.60	22.60	10
2002-03	Word	31.52	12.30	21		33.75	13.14	16

Appendix I: Comparison of SYTE and Comparison Group on 2005-06 PSSA Scores

Statistic	Say Yes	Control	Total
Mean	1222.4	1167.2	1204.6
N	44	21	65
SD	169.8	135.4	160.5
Mean	1104.3	1089.2	1099.5
N	44	21	65
SD	217.3	154.5	198.1
Mean	1191.8	1127.6	1173.8
N	36	14	50
SD	214.4	137.5	196.8

Appendix J: PSSA Achievement Levels

2005-06 PSSA Achievement Level By Subject and Experimental Group

		PSSA Math			PSSA Reading			PSSA Writing		
Achievement Level	Statistic	Say Yes	Control	Total	Say Yes	Control	Total	Say Yes	Control	Total
1	n	18	9	27	23	11	34	0	0	0
	% within group	40.9	42.9	41.5	52.3	52.4	52.3	0.0	0.0	0.0
2	n	13	9	22	10	7	17	20	11	31
	% within group	29.5	42.9	33.8	22.7	33.3	26.2	55.6	78.6	62.0
3	n	9	2	11	11	3	14	16	3	19
	% within group	20.5	9.5	16.9	25.0	14.3	21.5	44.4	21.4	38.0
4	n	4	1	5	0	21	21	0	0	0
	% within group	9.1	4.8	7.7	0.0	6.6	5.9	0.0	0.0	0.0
Total	n	44	21	65	44	21	65	36	14	50

Appendix K: Student Grades

Descriptive & Inferential Statistics for 2001-02 School Year

	Say Yes			Comparison			Total			
Course	Mean	N	SD	Mean	N	SD	Mean	N	SD	d
Data, Stats, & Probability	2.85	47	0.62	2.79	33	0.55	2.83	80	0.59	0.106
Number Systems	8.45	47	2.45	7.51	37	1.88	8.04	84	2.25	0.421
Other Curricular Area	10.53	47	2.27	11.76	37	8.74	11.07	84	6.03	-0.203
Patterns, Algebra & Functions	5.85	47	1.30	5.86	36	1.02	5.86	83	1.18	-0.008
Nature of Science	5.17	47	1.54	4.65	37	1.49	4.94	84	1.53	0.343
Social Studies	2.87	46	0.72	2.97	34	0.76	2.91	80	0.73	-0.137
Instructional Reading	9.49	47	3.23	9.39	36	3.35	9.45	83	3.26	0.031
Independent Reading	8.70	47	3.48	8.44	36	3.50	8.59	83	3.47	0.074
Stages of Writing	5.53	47	1.16	5.46	37	1.10	5.50	84	1.12	0.064
Social Skills Marks	14.00	47	0.00	17.03	37	18.41	15.33	84	12.22	-0.248
Work Habits Marks	28.00	47	0.00	31.41	37	20.71	29.50	84	13.75	-0.248

Appendix K (Continued): Student Grades

	Say Yes			Comparison			Total		
Course	Mean	N	SD	Mean	N	SD	Mean	N	SD
Data, Stats, & Probability	2.62	45	0.78	2.35	31	0.61	2.51	76	0.72
Geometry	2.29	45	0.82	2.48	31	0.63	2.37	76	0.75
Measurement	2.40	45	0.75	2.48	31	0.68	2.43	76	0.72
Number Systems	5.00	45	1.31	4.71	31	1.32	4.88	76	1.32
Other Curricular Area	52.91	44	3.83	56.40	30	13.64	54.32	74	9.25
Patterns, Algebra & Functions	2.16	45	0.77	2.29	31	0.53	2.21	76	0.68
Social Studies	2.76	45	0.61	2.58	31	0.50	2.68	76	0.57
Instructional Reading	11.58	43	2.99	10.46	28	2.43	11.14	71	2.81
Independent Reading	11.02	43	3.17	9.54	28	2.56	10.44	71	3.01
Stages of Writing	10.02	43	4.03	6.07	28	0.94	8.46	71	3.73
Science Courses	40.44	45	8.80	28.65	31	11.95	35.63	76	11.69
Writing Composition	29.05	43	24.33	29.36	28	23.53	29.17	71	23.85

	Say Yes			Comparison			Total		
Course	Mean	N	SD	Mean	N	SD	Mean	N	SD
Other Curricular Area	16.80	44	3.95	16.12	26	4.02	16.54	70	3.96
Independent Reading	12.09	44	2.78	12.16	25	2.13	12.12	69	2.55
Multiple Projects	13.00	1	.	13.63	8	0.52	13.56	9	0.53
Writing Comp	9.14	44	2.70	8.88	25	2.47	9.04	69	2.60
Service Learning Project	13.03	35	0.17	13.00	8	0.00	13.02	43	0.15

Appendix K (Continued): Student Grades

	Say Yes			Comparison			Total		
Course	Mean	N	SD	Mean	N	SD	Mean	N	SD
Other Curricular Area	10.57	42	3.61	11.15	20	4.30	10.76	62	3.82
Writing Comp	2.67	6	2.66	2.00	7	2.65	2.31	13	2.56

	Say Yes			Comparison			Total		
Course	Mean	N	SD	Mean	N	SD	Mean	N	SD
Other Curricular Area Course	11.09	22	3.64	8.53	15	2.64	10.05	37	3.47

Acknowledgements

The authors would like to thank Dr. Michael Schlesinger and Dr. Daniel Pietrowski of the School District's Office of Research & Evaluation. They provided all the data necessary for this report and were available throughout the analysis to answer numerous questions about the data. In addition, we wish to thank David Osher and his colleagues at the American Institutes for Research for their useful peer review of the report. Dr. Ruth Curran Neild, University of Pennsylvania, provided a thorough and thoughtful review of a draft of the report. Dr. Robert Boruch, University of Pennsylvania, provided advice and guidance at several points in the process.

We also appreciate the work of our colleagues at Research for Action who took time to review and provide feedback on drafts of the report. Gail Danckert, Judy Adamson, Cecily Mitchell, and Morgan Riffer all provided feedback and support in producing this report.

Authors

Herbert Turner, III, Ph.D. is a graduate of the University of Pennsylvania's *Policy Research, Evaluation, and Measurement Program.* Dr. Turner is the founder and president of ANALYTICA, Inc. He has 25 years of quantitative research experience in the private, public, and education sectors. He has collaborated on a number of important research projects with professionals in prominent research organizations in the United States including the American Institutes for Research (AIR), The U.S. Department of Education What Works Clearinghouse, The Campbell Collaboration, The Cochrane Collaboration, the Educational Testing Service (ETS), Mathematica Policy Research, and the Consortium for Policy Research in Education (CPRE) at PENN. Dr. Turner has published a number of articles on randomized controlled trials (RCT) and on synthesis of studies that used a RCT design, with leading researchers and scholars in the field. He also lectures on research methods, statistical analysis, and systematic reviews at the University of Pennsylvania and the University of Central Florida.

Jason Schoeneberger
Jason Schoeneberger has over seven years of experience in data processing and analysis. Primarily focused in the public education arena, Jason has designed and carried out numerous research and evaluation activities to assist large urban school districts in determining 'what works', including Broward County Schools in Ft. Lauderdale, Philadelphia schools in Pennsylvania, and Charlotte-Mecklenburg Schools in North Carolina. Jason has highly developed skills in the manipulation and creation of large data files using various statistical packages. In addition, Jason is familiar with a wide-array of analytical tools and methods to conduct any number of applied research projects. He also has worked in the high-stakes testing sector helping to create the scoring routine for the recently developed Step 2-Clinical Skills test for the National Board of Medical Examiners. Jason also has experience in the conduct and analysis of web-based surveys, and has facilitated focus groups charged with deliberating legal outcomes for trial attorneys in Florida.

Tracey Hartmann, Senior Research Associate, works on a variety of evaluation and policy research projects at RFA. She is currently leading a teams evaluating the School District of Philadelphia's pilot of Parent Leadership Academies and has worked on studies including an evaluation of the federally funded college readiness program, GEAR UP, an evaluation of Say Yes to Education, a scholarship guarantee program and Going Small, a policy study which looks at the development of public-private partnerships to run small high schools in Philadelphia. Tracey has worked in a variety of educational and youth serving settings and has held research positions in several Philadelphia organizations since 1997, most recently as Research Associate at Public/Private Ventures. She earned her doctoral degree in Human Development from the University of Pennsylvania in 2004.

Eva Gold, Ph.D. is a Founder of Research for Action and a research director of the *Learning from Philadelphia's School Reform* project. She has served as primary investigator for numerous local and national studies examining the dynamics of parent, community, school relations. She (and Elaine Simon) led the research for the Education Organizing Indicators Project, a joint project of Research for Action and the Cross City Campaign for Urban School Reform. This project culminated in the report series *Strong Neighborhoods, Strong Schools* which presents a process for documenting the contributions of parent/community organizing groups to strengthening communities and improving schools. Most recently, Dr. Gold was the principal investigator of the Say Yes to Education program, a scholarship guarantee program, which incorporates a significant parental engagement component. In addition to her interest in parent, community, and school dynamics, Dr. Gold's research interests include home and school literacies, and the politics of urban education. She is a Guest Lecturer in the Urban Studies Program and Graduate School of Education at the University of Pennsylvania, where she teaches courses on community activism and school reform and methods of data analysis and reporting.