# Analyzing the Effects of Daily Physical Education in Middle Schools on Obesity: Evidence from Pennsylvania's Active Schools Program

Stephen Erfle, Ph.D.
International Business and Management
Dickinson College
Carlisle, PA 17013
erfle@dickinson.edu

**ABSTRACT**

**Background:** The Pennsylvania Department of Health (PADoH) launched the Active Schools Program (ASP) to encourage daily physical activity. ASP schools instituted a minimum of 30 minutes of daily physical education (PE). Control schools maintained their standard schedule of non-daily PE. Both administered a fitness assessment at the beginning and the end of the school year.

**Objectives:** Do students at schools with daily PE have significantly better outcomes than students at schools without daily PE? We examine performance at the midstream Behaviors level and at the downstream Health outcomes level. Behavioral outcome measures include changes in curl-ups, push-ups and mile run. Health outcome measures include change in BMI and change in BMI percentile (*dB%*) which adjusts for age and sex differences.

**Method:** Thirty ASP schools and nine control schools provided complete pre- and post-assessments ($N_{\mathrm{ASP}}$ = 6,693, $N_{\mathrm{control}}$ = 3513). ASP schools were allowed to purchase commercial programs from a list provided by PADoH or they were allowed to propose their own program. This led to five ASP subsamples: three multiple schools subsamples (HOPSports® [9 schools, 2,066 students], SPARK™ [7 schools, 1,069 students], and CATCH® [2 schools, 601 students]); “Other” aggregates together schools which chose HEALTHY PE, Physical Best, and Project Fit America® (331 students); and “Own” includes schools that created their own program by purchasing fitness equipment or creating a fitness course or a walking trail (9 schools, 2,626 students). Statistical analysis employs difference between means tests and regression analysis. Treatment effect size is measured using standardized mean difference (SMD).

**Results:** The ASP was successful in altering obesity outcomes in middle school students relative to students at control schools. The most successful individual programs studied were HOPSports® and SPARK™. CATCH® and Other schools had more ambiguous overall effects. Own schools did significantly worse on dB% but better on each behavioral outcome than control schools. In general, health SMDs were more modest than behavioral SMDs but significant health outcomes were present, especially for HOPSports® and SPARK™.

In general, females had a relatively higher treatment effect on health outcomes and males had a relatively higher treatment effect on behavioral outcomes at schools with formal programs. These effects are most equally balanced at HOPSports®. SPARK™ was the only program studied in which all five outcome measures favored males.

For at-risk students (B% ≥ 85), regression analysis of dB% as a function of starting B%, behavioral outcome measures, sex, and program suggests that much of the health outcome benefit is due to differences in behavioral outcomes achieved by individuals involved in daily PE. A significant program effect remains for HOPSports® and CATCH® but not SPARK™ once one controls for behavioral outcomes. Change in push-ups and change in mile run were significant predictors for all program subsamples, but change in curl-ups was not statistically significant for the HOPSports®, SPARK™, and Other subsamples.

**Conclusion:** Students exhibit positive health benefit from access to daily PE relative to students without daily PE. Those benefits are greater for midstream behavioral outcomes than downstream health outcomes.

## INTRODUCTION

The dramatic increase in pediatric obesity since the 1970s coincides with major changes in children's food and activity environments (Balistreri & Hook, 2011; Morrill & Chinn, 2004). Calorically dense, inexpensive, fat- and sugar-laden foods and drinks have become increasingly available and options for sedentary behavior have ballooned and so have children (Han & Powell, 2013; Piernas & Popkin, 2011; Pérez-Escamilla et al., 2012). Laws attempting to increase academic performance such as the *No Child Left Behind Act* have created increased pressure within schools to reduce the time devoted to physical education and recess (Public Law 107-110, 2002). These are ominous trends not only for children's health, but also for their academic performance because research shows that obesity is positively related to physical inactivity, and physical activity is positively related to academic performance (Du Toit, Pienaar, & Truter, 2011; Florin, Shults, & Stettler, 2011; Roberts, Freed, & McCarthy, 2010; Shephard, 1997; Telford et al., 2012).

Despite national recommendations that all school-age children receive daily physical education (*PE*) (Heidorn, 2011), inclusion of daily PE remains elusive in many states. "It does not appear that physical education is a very good fit within educational programs in the current political, economic, and educational environment" (Ennis, 2006, p. 54). In May 2010, Pennsylvania's State Board of Education proposed a sweeping set of reforms to school nutrition standards and physical activity and PE requirements that would position the state as a national leader in school wellness policies. A seminal feature of these reforms was the introduction in Section 12.84 of a minimum of 30 minutes of daily, moderate to vigorous physical activity and in Section 12.85 a requirement of 150 minutes of PE per week in elementary school and 225 minutes per week in middle and secondary school (State Board of Education, Pennsylvania, May 2010). By November 2010, the proposed Section 12.85 PE standards had been set aside. “In response to feedback from education stakeholders, **the Board will remove the physical education requirements from this rulemaking in the final version**” (Torsella, 2010, p. 1, emphasis included in the original).

Increased school-based physical activity should be associated with decreased obesity. Unfortunately, this seemingly simple proposition has had only modest empirical support according to recent meta-analyses of the literature (Harris, Kuramoto, Schulzer, & Retallack, 2009; Katz, O'Connell, Njike, Yeh, & Nawaz, 2008; Shaya, Flores, Gbarayor, & Wang, 2008). The present research examines whether daily PE in schools has a statistically significant impact on obesity relative to non-daily PE that is currently the norm due to competing pressures including the *No Child Left Behind Act* (Katz, 2009). This chapter examines this issue at both the midstream Behaviors level and at the downstream Health outcomes level using the causal pathway that influences health and weight outcomes logic model of obesity (Katz, 2009).

## METHOD

### Participants

Pennsylvania Department of Health (*PADoH*) launched the Active Schools Program (*ASP*) to encourage daily physical activity in middle schools across the Commonwealth. This program provided schools with a one-time $15,000 grant in the form of a $5,000 Preventive Health and Health Services Block Grant from the Centers for Disease Control (*CDC*) distributed by PADoH together with a 2 for 1 matching grant from various partner organizations arranged by PADoH. In exchange, ASP schools agreed to undertake a regiment of 30 minutes of daily PE and to assess students on physical fitness (*PF*) performance and height and weight at the start and the end of the 2009 – 2010 school year using the ASP protocol.

The Robert Wood Johnson Foundation supported this research by funding a control school analysis through Active Living Research Rapid Response grant # 68311 with the understanding that control schools would undertake the ASP assessments at the start and the end of the 2010 – 2011 school year but would otherwise maintain their standard schedule of non-daily PE. Control schools were provided with a $700 grant per assessment for administering the ASP protocol to their students. This protocol has received IRB approval.

This analysis uses the combined dataset. A total of 14,603 students were assessed twice; 9,722 from 30 ASP schools and 4,881 from nine control schools. The participants were evenly distributed across sex with 49.5% of ASP students, and 50.7% of control students, being female. These students were from urban, suburban, and rural schools spanning the Commonwealth of Pennsylvania.

The PADoH prioritized the introduction of ASP in school districts with above average rates of students who were at-risk for obesity by having a body mass index percentile (*B%*) of 85 or above and who are hereafter described as *At-risk* according to CDC guidelines (Ogden & Flegal, 2010). Control schools were chosen with similar priorities in mind. Summary statistics by school district are aggregated to two grade groupings, grades K - 6 and 7 - 12 by PADoH. An exact comparison of At-risk levels is not possible since the ASP targets grades 6 - 8. During the 2008 – 2009 school year, the most recent for which data was available from PADoH, 31.5% of the 979,048 grades K - 6 and 33.5% of the 839,298 grades 7 - 12 students are At-risk according to CDC guidelines (Department of Health, Pennsylvania, 2011). The proportions of At-risk students in ASP and control schools exceed that of the entire Commonwealth. The fall assessment had 37.7% of ASP students and 37.0% of control students identified as At-risk. By the spring assessment, these percentages had declined to 35.9% and 36.8% at ASP and control schools, respectively. It is worth noting that these percentages presage the final result from this analysis. Schools that had daily PE had a decline in At-risk students relative to those that did not have daily PE.

Schools participating in the ASP were allowed to purchase commercial programs from a list provided by PADoH or they were allowed to propose and implement their own program. This led to five ASP subsamples. Three subsamples are for programs chosen by multiple ASP schools (*S*): HOPSports® (labeled HOPS in tables and figure, S = 9); SPARK™ (S = 7); and CATCH® (S = 2). The subsample labeled *Other* (S = 3) aggregates together schools which chose

HEALTHY PE, Physical Best, and Project Fit America®. Finally, *Own* (S = 9) includes schools that used the ASP funds to create their own program by purchasing fitness equipment or creating an outdoor fitness course or creating a walking trail.

**Procedure**

PADoH provided assessment protocols to ASP instructors. School representatives were required to participate in a webinar on assessment protocols and use of the reporting template to ensure minimal bias in implementation across schools. The assessment protocols were also sent to schools and published on the ASP website. School nurses, who are PADoH employees, measured height and weight using established PADoH protocols (Department of Health, Pennsylvania, 2013). The ASP assessment included the mile run, curl-ups, push-ups, sex, age, grade, height, and weight. In the curl-up test, students had 60 seconds to perform as many repetitions as possible, the same time limit used by the *President's Challenge*. For the push-up test, students were instructed to do push-ups until failure. Teachers and school nurses entered the data into a modified version of an Excel file created by the CDC for use in schools (Centers for Disease Control and Prevention, 2009). This analysis is restricted to the 10,206 students having complete data on all variables for both assessments; 6,693 from ASP schools and 3,513 from control schools. Slightly more females had missing data than males, leading to an ending sex balance of 49.2% females at ASP and 50.4% females at control schools.

Excel was used for data cleaning and SPSS was used for statistical tests. These tests allow the reader to examine performance differences for more and more finely disaggregated groups. Tests include independent sample *t* tests, one-way analysis of variance tests paired with the LSD method of multiple comparisons, and multiple linear regression. Standardized mean difference (SMD) was used to examine effect size. This chapter uses $p = 0.05$ to test for statistical significance.

**Five outcome measures viewed two ways.** The three ASP behavioral outcomes measure change in PF performance levels, (*dPF*). Three dPFs include changes in curl-ups, push-ups and the mile run (denoted *dC*, *dP*, and *dM*). Health outcome measures include change in body mass index (*dB*) and change in B% (*dB%*) which adjusts for differences in age and sex that may confound BMI comparisons. These five measures are correlated to one another. Each of the ten correlations between these five measures are significant at the $p < .001$ level, the smallest correlation is between dC and dM at $r = -.039$. The highest correlation among dPFs is between dC and dP at $r = .116$. The correlation between the two health measures is much higher at $r = .738$ and the highest correlation between health and behavioral measure is between dB and dM at $r = .141$.

Two versions of each outcome measure are examined because schools differed with regard to *dDays*, the number of days between the fall and spring assessments. In this instance, part of the difference in outcome measure may be due to differences in the length of time between assessments. A straightforward method to control for this difference is to create an annualized version of each outcome measure by multiplying the outcome measure by 365/dDays.

## RESULTS

Summary information for five outcome measures is provided in Table 1 for various subsamples. A quick comparison of mean values from control and ASP subsamples confirms that ASP schools have superior performance to control schools on all five outcome measures, four of which are statistically significant differences (as will be seen in Table 3). More interesting than this however, are the patterns that emerge when we subdivide the ASP data by program used at each ASP school.

***** TABLE 1 ABOUT HERE *****

Table 2 provides summary information for dDays by subsample. Subsample dDays averages range from a low of 206.8 for control schools to 265 for the two schools in the CATCH® subsample. Optimally, there should be no significant dDays differences between schools. In fact, an LSD *t* test shows that all but one of the 15 pairwise comparisons have significantly different dDays means (SPARK™ and Other, $p = .552$), 13 of which are significantly different at the $p < .001$ level. Differences exist for a variety of factors, even at ASP schools. At control schools, the need to work assessments into the school day at schools without the benefit of daily PE classes together with the later starting time that occurred at a number of control schools at the beginning of the 2010 – 2011 school year led to control assessments being about one month shorter, on average, than assessments at ASP schools (206.8 days versus 235.7 days).

***** TABLE 2 ABOUT HERE *****

**Analyzing program differences**. Table 3 provides difference between means tests comparing the control school mean with the ASP full sample as well as the five ASP subsamples for each of the five outcome measures for both nominal and annualized differences. A comparison of *p* levels in the two versions shows that only two of the statistically significant nominal results from Table 3 fail to remain statistically significant in annualized form: Own dB being less healthy than control schools (nominal $p = .005$, annualized $p = .256$) and CATCH® dB% being healthier than control schools (nominal $p = .039$, annualized $p = .084$). Given these differences, the more conservative approach is to conclude that annualized outcome measures provide a more accurate description of the net benefit of the ASP since they control for outcome effect differences that may be due to time between measurements differences.

***** TABLE 3 ABOUT HERE *****

Both nominal and annualized outcomes show consistent patterns that suggest a clear dominance by HOPSports® and SPARK™. Schools using CATCH® and Other programs exhibit more mixed results with fewer statistically significant positive health and behavioral outcomes. Interestingly, schools creating their own program did worse according to the health outcome measures but better according to the behavioral outcome measures. Given these differences between programs, further analysis will examine ASP subsamples relative to control schools rather than treat all ASP schools as part of a unified whole. Nonetheless, the full ASP sample results are provided for comparison.

**Standardized mean differences**. In order to create a uniform scale of effect size across annualized outcome measures and programs, SMD (SMD = mean difference/total standard deviation) and 95% confidence interval (CI) on SMD is calculated. A statistically significant SMD is one that has upper and lower 95% CI bounds of the same sign and that therefore the 95% CI does not contain 0. To simplify the visual exposition in Figure 1, each SMD is calculated so that a positive difference represents a positive health or behavioral outcome for the program relative to the control and a negative difference represents the reverse situation. This requires that the push-up and the curl-up annualized mean difference numerators from Table 3 be multiplied times -1 in calculating SMDs. One final SMD, the start of year average B% status as measured by mean B% in the Fall (*B%Fall*) is included to the left of each subsample for comparison purposes in Figure 1. Subsample B%Fall means range from 65.1 for HOPSports® to 70.2 for CATCH®, with B%Fall differences calculated as the control sample B%Fall mean of 66.4 minus Program mean B%Fall. A quick scan of B%Fall SMDs in Figure 1 shows that only the CATCH® subsample had a significantly higher mean B%Fall than the control sample; the other subsamples did not have significantly different means from the control sample. It should be noted that the HOPSports® subsample had significantly lower mean B%Fall than all subsamples except the control subsample. Nonetheless, the percentage of HOPSports®

students with B%Fall in the At-risk range, 34.7%, still exceeds the Commonwealth-wide At-risk percentages outlined above.

***** FIGURE 1 ABOUT HERE *****

The full ASP subsample outcomes provide, as expected, a blended version of the general contours of individual programs. Among health outcome measures, dB exhibits a more significant effect than dB% but both have smaller effect size than curl-ups and mile run. All three behavioral measures exhibit significant differences but push-ups exhibit a substantially smaller effect size than the other two dPFs.

HOPSports® and SPARK™ appear to have roughly the same overall effectiveness when viewed across outcome measures. HOPSports® has the edge on behavioral outcomes while SPARK™ holds a modest edge on health outcomes. CATCH® shows some promise with one health and one behavioral outcome significantly better but the other health outcome is not significantly better and, on the other two behavioral metrics, CATCH® is significantly worse than control schools. These are the only two behavioral metric comparisons for which daily PE exhibits a significantly inferior outcome to non-daily PE. It should be acknowledged that the CATCH® subsample has only 601 students from two schools and it is therefore not as robust of a subsample as HOPSports® or SPARK™. Other programs had two significant behavioral outcomes but they were more modest in magnitude than those produced by HOPSports® or SPARK™. Finally, Own showed significant positive behavioral outcomes on all three measures but mean dB% was significantly worse than control schools.

**Sex differences**. Another issue worth examining is whether substantive sex differences exist with regards to the programs chosen by ASP schools. Table 4 provides mean annualized outcome and SMD with 95% CI between control and program for each of the five ASP subsamples for each measure. As was done in Figure 1, SMDs are defined so that positive differences indicate higher mean performance by program than control schools. Also included at the bottom of Table 4 are aggregated measures by program and by outcome and mean annualized outcomes at the control schools by sex.

***** TABLE 4 ABOUT HERE *****

Each of the SMD outcomes for specific programs shown graphically in Figure 1 is disaggregated into female and male components in Table 4. Consider, for example, the HOPSports® dB%/Yr SMD of 0.10, 95% CI [0.05, 0.15] in Figure 1. This significant outcome difference is due to stronger dB%/Yr performance by females than males. This conclusion can be reached by comparing the female and male data for this outcome in Table 4. The female average dB%/Yr of -2.40 is far stronger than the female control mean of 0.79. HOPSports® males do have healthier mean dB%/Yr performance, -2.34, than control males, -1.93, but the difference is not significant. These differences translate into a female dB%/Yr SMD of 0.19, CI [0.11, 0.26], and the male SMD of 0.02, CI [-0.05, 0.10] in Table 4. HOPSports® females have dB%/Yr that are significantly better than control females, but HOPSports® males do not show a significant improvement on this outcome measure relative to control males.

Of particular interest are the three statistically significant negative SMD outcomes in Figure 1. These outcomes are CATCH®'s dP/Yr and dM/Yr and Own's dB%/Yr. Two of these are based on balanced inferior performance by both sexes but the third is based on an asymmetry between females and males. The CATCH® dM/Yr SMD of -0.10 is based on a female SMD of -0.25 and a male SMD of 0.05. Put in terms of average dM/Yr, Table 3 shows that CATCH® had an average dM/Yr that was, on average, 0.41 minutes slower than control schools but Table 4 shows that females were more than a minute slower than control schools (-1.01 = -0.03 - 0.98) and males are 0.21 minutes faster [0.21 = 0.01 – (-0.20)].

The male versus female SMD comparisons suggest that, in general, females have a relatively higher treatment effect than males on health outcomes and males have a relatively higher treatment effect than females on behavioral outcomes at schools where formal programs were chosen. These effects are most equally balanced at HOPSports®. SPARK™ was the only program studied in which all five outcome measures favor the same sex. SPARK™ males have, on average, a 0.12 SMD advantage over females (0.12 = 0.30 - 0.18). Interestingly, SPARK™ is the only program where males exhibited significant positive results on dB%/Yr; by this same outcome measure, females had significant positive results at HOPSports®, SPARK™, and CATCH®. The schools which chose to create their own program did not show the same health outcomes versus behavioral outcomes differences across sexes. At these schools, all three behavioral outcomes show significant positive effects but the effect size for curl-ups favors females and the mile run favors males.

**Regression analysis of dB% for At-risk students**. The pediatric obesity epidemic is most pressing for those who are most at-risk of being overweight or obese. The CDC has recommended using a cutoff B% of 5.0 for underweight, 85.0 for overweight, and 95.0 for obese (Centers for Disease Control and Prevention, 2009). The full sample had 334 students who had B% < 5.0 for at least one of the assessments, and who are therefore considered underweight, 5,685 who were in the [5.0, 85.0) normal B% range for both assessments, and the remaining 4,187 who were in the [85.0, 100] At-risk B% range for at least one of the assessments. A negative dB% is unhealthy for those in the underweight range, ambiguous for those in the normal range, and healthy for those in the At-risk range. The regression analysis will focus on the At-risk 41.0% of the full sample because health implications of positive or negative changes in B% can be unambiguously delineated for these At-risk students.

A difference between means test shows that the percentage of At-risk ASP students, 41.2%, does not differ significantly from the percentage of At-risk control students, 40.7%, $t = 0.48$ and $p = .64$. Because it allows the most direct comparison across sex and grade from schools with disparate time between assessments, dB%/Yr is used as the dependent variable. Alternative specifications, not reported here, using non-annualized versions of the dependent and independent variables and using different B% subsample cutoff points (such as 50 rather than 85) show little difference in terms of sign and significance of the coefficients for dPF, sex, and program.

Students with low B%s tend to have positive dB% and students with high B%s tend to have negative dB%: dB% is a nonlinear function of starting B%. As a result, B%Fall is included in cubic functional form in each model in Table 5. Six At-risk subsamples are examined: ASP and five ASP subsamples. Each includes At-risk control students, therefore, each is denoted with a + sign after its name in Table 5. Two models, I and II, are provided for each subsample. The first controls for sex and starting B%. The second includes dPF covariates divided by 10 so that dPF coefficients are increased by a factor of 10.

All models include a program dummy variable taking the value of 1 for students in the program group and 0 for students in the control group. The program dummy is of central importance because it provides a best guess estimate of the effect of the program on B% for At-risk individuals, all else held constant. The bottom of Table 5 includes the results of the Program Net Effect calculations on dB%/Yr by sex using model II evaluated at At-risk program and control mean dPF values (from Table 6).

***** TABLE 5 ABOUT HERE *****

Comparing model I with II for each subsample shows that, as expected, a positive dPF is predictive of decreased body mass. Each change in activity is of the expected sign and the change in adjusted $R^2$ from including dPF variables range from 1.0% to 1.5%. Push-ups and mile run are significant predictors in all subsamples (and most are significant at the $p < .001$ level) but curl-ups fails to achieve significance in the HOPS+, SPARK+, and Other+ subsamples.

Much of the effect, in terms of reduced B%, that is noted in the model I program coefficient is due to increased PF for students in the program as can be seen from its coefficient in model II. With the exception of CATCH®, the difference between the model I program coefficient and its model II counterpart is always a negative number. The most dramatic of these differences is -0.97 for HOPSports®. The CATCH® difference is due to the substandard dPF performance by CATCH® relative to control students for two of three activities in Figure 1.

The *Program Net Effect* for each sex is an estimate of dB%/Yr that takes into consideration expected changes in behavioral outcome differences due to the program. It is calculated using model II by summing the program and male (as appropriate) coefficients together with the dPF slope coefficient times the difference between program and control At-risk mean dPF values summed across activities. For example, the HOPSports® female Program Net Effect of – 2.44 is calculated as:

$$-2.44 = -1.58 - 0.19\cdot(11.3 - 0.67)/10 - 0.79\cdot(3.98 - 2.42)/10 + 2.57\cdot(-1.99 - 0.10)/10,$$

based on the At-risk female HOPSports® and control subsample dPF means shown in Table 6. This estimate is close to the female HOPSports® program estimate of -2.55 in model I. Similarly, the HOPSports® male Program Net Effect calculated estimate of -2.35 using model II (and using At-risk female HOPSports® and control subsample dPF means) is close to the model I male program estimate of -2.49 (-2.49 = -2.55 + 0.06).

The sex differences across programs noted in Table 4 are confirmed in the regression analysis in Table 5. None of the male dummy variables in Table 5 are significantly different from zero. However, in each case, the model II male coefficient is larger than its model I counterpart due to differences in PF performance by sex. Differences in Program Net Effect across sex are small except at SPARK™ and CATCH®, where the Program Net Effect is more than 75% larger for males than females [94% = (2.20 – 1.13)/1.13 for SPARK™ and 76% = (2.60 – 1.48)/1.48 for CATCH®].

**Behavioral outcomes for At-risk students.** One of the main benefits of daily PE is the benefit that accrues to behavioral outcomes, as measured by dPF. This was seen (a) in Figure 1 for the full sample, (b) in the SMD by sex disaggregation in Table 4, and (c) in the Program Net Effects in Table 5. It is confirmed in Table 6, which disaggregates the behavioral outcomes in Table 4 by At-risk status. Table 6 provides 24 subsample mean values for each of the three annualized behavioral outcome measures: 5 ASP subsamples plus control × Sex × At-risk status.

***** TABLE 6 ABOUT HERE *****

Table 6 also provides SMD and 95% confidence interval for each behavioral measure by Program × Sex × At-risk status. These data confirm that ASP students have superior performance relative to control students, especially with curl ups and mile. All 20 curl-up SMDs are positive and 19 are significantly different from control. For the mile run, 18 of 20 have positive SMDs and 15 of these are significantly different from control. The remaining two are CATCH® female subsamples and for one of these, the At-risk subsample, mile is significantly worse than control. For push-ups, the results are more balanced; six of 11 positive coefficients are significant, while four of the nine negative coefficients are significant. Three of the four significant negative push-up coefficients are for Not At-risk subsamples; CATCH® female is the sole significant negative At-risk push-up result.

One would hope that programs targeting obesity would have the most substantive effect on the At-risk subsample. There are 30 possible comparisons of At-risk with Not At-risk for ASP schools: 5 Programs × 3 Behavioral outcomes × Sex. Two comparisons are possible.

The mean increase in performance is likely to be smaller for the At-risk subsample than for the Not At-risk subsample. All six of the control school comparisons show superior performance increases by students in the Not At-risk subsample, but for only one of these, female push-ups, was the difference significant. In the ASP program subsamples, 11 of the 30 comparisons have At-risk mean outcomes that exceeded the Not At-risk mean outcomes but none of these differences were significant. Four of the 19 comparisons where the Not At-risk mean outcome exceeds the At-risk mean outcome are significant. These four, where At-risk females did significantly worse than Not At-risk females, were HOPSports® push-ups, CATCH® mile run, and Own curl-ups and mile run.

When one considers how each group did relative to their control counterpart, a more positive pattern emerges. Comparing SMDs between At-risk and Not At-risk subsamples incorporates how students of different At-risk statuses performed in the absence of the program. By this criterion, the programs proved to be especially effective for the At-risk subsample; 20 of the 30 SMDs are higher for the At-risk than the Not At-risk counterpart. This can also be seen using the aggregated SMDs at the bottom of Table 6. By this criterion, CATCH® females are the only subsample in which Not At-risk students performed better than At-risk students; as noted above, they did so by doing worse than control students on push-ups and the mile run.

## CONCLUSION

The ASP was successful in altering middle school student obesity outcomes at ASP schools relative to students at a group of control schools. The ASP showed that students were able to improve behavioral outcomes such as increasing PF performance and improve health outcomes such as decreasing BMI and B% by incorporating daily PE into the school day relative to schools that did not have daily PE. Difference between means tests and analyzing effect size using SMDs suggest that while there were significant health outcomes from the ASP, the behavioral outcomes were more dominant than the health outcomes.

Regression analysis suggests that a significant fraction of the downstream health outcome is, indeed, due to midstream behavioral outcomes. These results conform to outcomes documented by other researchers: midstream outcomes are more likely to occur than downstream outcomes. According to Katz (2009): "Most interventions are apt to influence upstream or midstream variables; only an aggregation of effective programming is likely to produce meaningful change in the downstream variables" (p. 262).

The ASP allowed schools to choose from a variety of alternative commercial programs or to propose and create their own program using the one-time ASP grant. The schools chose six commercial programs but only three were used by more than one school, HOPSports®, SPARK™, and CATCH®. Nine schools chose to create their own program. As a result, a separate analysis was performed on five ASP subsamples. The analysis is further disaggregated in order to examine differences that may exist by sex and At-risk status.

HOPSports® and SPARK™ appear to have the greatest overall effect, with CATCH®, Other programs, and Own program having more mixed results. In comparing across outcomes using SMDs, HOPSports® appears to have the edge on behavioral outcomes and SPARK™ has the edge on health outcomes. With the exception of CATCH®, at least two of the three behavioral outcome SMDs for each program show statistically significant benefit relative to control schools. Mile run and curl-ups exhibit the most robust results; with push-ups the results are more mixed.

HOPSports® appears to be more balanced across sexes while SPARK™ (and CATCH®) provides greater benefits to males than females. Sex differences in SMD also vary by outcome measure. Females show greater effect on dB% but males show greater effect on push-ups and mile run.

Comparisons of behavioral outcome SMD by At-risk status show that, with the exception of CATCH® females, those who are At-risk for obesity have a higher average SMD across behavioral outcomes than those who are Not At-risk.

Put another way, the harm that accrues from not having daily PE accrues more heavily to those that are at-risk for obesity.

One of the main benefits of the ASP, regardless of whether the school chose to use the ASP funds to purchase an established physical activity program or created their own program, is increased dPF performance of ASP students relative to control students. In each of the subsamples except CATCH®, the regression model that includes dPF covariates shows a smaller program coefficient than the model that excludes dPF covariates. Indeed, the statistically significant program coefficient in the SPARK™ model fails to remain statistically significant once dPF covariates are introduced into the model.

The data also shows that there are significant benefits to be obtained from the dedicated physical activity programs relative to self-created programs. Two of the programs, CATCH® and SPARK™, include nutrition information along with their physical activity program. This may well explain some of the strength of CATCH® once we control for changes in PF performance. Of course, part of this is also due to the weak behavioral outcomes exhibited by CATCH® students, especially by At-risk females, mentioned above.

In net, daily PE is a statistically significant weapon in combating the pediatric obesity epidemic. Unsurprisingly, it has a greater effect on behavioral outcomes than health outcomes. The impact is also greater for At-risk students than those who are Not At-risk for obesity. Among programs studied, the HOPSports® program demonstrated the strongest and most balanced impact across sex and At-risk status categories.

Table 1. *Summary Statistics for Five Outcomes, dx, Measured as dx = Spring x - Fall x.*

| Outcome Measure, dx | | | Subsample | Nominal Outcome | | | Annualized Outcome | | |
|---|---|---|---|---|---|---|---|---|---|
| | | | | Mean | Std. Dev. | Std. Error | Mean | Std. Dev. | Std. Error |
| Health Outcome Measures | ΔBMI (dB) | | Control (S=9) | **0.37** | 1.47 | 0.025 | **0.68** | 2.58 | 0.044 |
| | | | ASP (S=30) | **0.29** | 1.56 | 0.019 | **0.45** | 2.39 | 0.029 |
| | | ASP subsample | HOPS (S=9) | **0.17** | 1.58 | 0.035 | **0.25** | 2.42 | 0.053 |
| | | | Spark (S=7) | **0.10** | 1.44 | 0.044 | **0.15** | 2.32 | 0.071 |
| | | | Catch (S=2) | **0.19** | 1.77 | 0.072 | **0.26** | 2.45 | 0.100 |
| | | | Other (S=3) | **0.37** | 1.15 | 0.063 | **0.60** | 1.82 | 0.100 |
| | | | Own (S=9) | **0.48** | 1.57 | 0.031 | **0.75** | 2.42 | 0.047 |
| | | | Total (S=39) | **0.32** | 1.53 | 0.015 | **0.53** | 2.46 | 0.024 |
| | ΔBMI%ile (dB%) | | Control (S=9) | **-0.37** | 10.75 | 0.181 | -0.56 | 18.6 | 0.314 |
| | | | ASP (S=30) | **-0.79** | 11.39 | 0.139 | **-1.26** | 17.6 | 0.215 |
| | | ASP subsample | HOPS (S=9) | **-1.45** | 12.09 | 0.266 | **-2.37** | 18.7 | 0.412 |
| | | | Spark (S=7) | **-2.36** | 10.40 | 0.318 | **-3.76** | 16.7 | 0.510 |
| | | | Catch (S=2) | **-1.39** | 12.37 | 0.504 | **-1.92** | 17.0 | 0.695 |
| | | | Other (S=3) | 0.45 | 7.47 | 0.410 | 0.77 | 11.9 | 0.653 |
| | | | Own (S=9) | 0.35 | 11.26 | 0.220 | 0.52 | 17.5 | 0.342 |
| | | | Total (S=39) | **-0.65** | 11.17 | 0.111 | **-1.02** | 17.9 | 0.178 |
| Behaviorial Outcome Measures | ΔCurl-ups (dC) | | Control (S=9) | **1.27** | 12.21 | 0.206 | **2.05** | 22.31 | 0.376 |
| | | | ASP (S=30) | **6.35** | 14.57 | 0.178 | **9.83** | 22.25 | 0.272 |
| | | ASP subsample | HOPS (S=9) | **8.18** | 14.45 | 0.318 | **12.86** | 22.56 | 0.496 |
| | | | Spark (S=7) | **8.16** | 15.44 | 0.472 | **12.80** | 24.30 | 0.743 |
| | | | Catch (S=2) | **9.20** | 22.43 | 0.915 | **12.79** | 31.05 | 1.267 |
| | | | Other (S=3) | **6.07** | 7.78 | 0.428 | **9.66** | 12.37 | 0.680 |
| | | | Own (S=9) | **3.55** | 11.98 | 0.234 | **5.57** | 18.74 | 0.366 |
| | | | Total (S=39) | **4.60** | 14.01 | 0.139 | **7.15** | 22.57 | 0.223 |
| | ΔPush-ups (dP) | | Control (S=9) | **2.33** | 9.29 | 0.157 | **4.30** | 16.50 | 0.278 |
| | | | ASP (S=30) | **3.38** | 8.88 | 0.109 | **5.32** | 13.79 | 0.169 |
| | | ASP subsample | HOPS (S=9) | **4.09** | 8.60 | 0.189 | **6.47** | 13.65 | 0.300 |
| | | | Spark (S=7) | **2.41** | 9.41 | 0.288 | **3.92** | 14.85 | 0.454 |
| | | | Catch (S=2) | 0.45 | 10.10 | 0.412 | 0.62 | 13.86 | 0.565 |
| | | | Other (S=3) | **2.53** | 6.13 | 0.337 | **4.01** | 9.71 | 0.534 |
| | | | Own (S=9) | **3.99** | 8.71 | 0.170 | **6.22** | 13.59 | 0.265 |
| | | | Total (S=39) | **3.02** | 9.04 | 0.089 | **4.97** | 14.78 | 0.146 |
| | ΔMile time (dM) | | Control (S=9) | -0.01 | 2.16 | 0.036 | -0.01 | 3.94 | 0.067 |
| | | | ASP (S=30) | **-0.79** | 2.44 | 0.030 | **-1.25** | 3.79 | 0.046 |
| | | ASP subsample | HOPS (S=9) | **-1.17** | 2.19 | 0.048 | **-1.84** | 3.43 | 0.076 |
| | | | Spark (S=7) | **-0.89** | 2.40 | 0.073 | **-1.44** | 3.88 | 0.119 |
| | | | Catch (S=2) | **0.28** | 3.12 | 0.127 | **0.40** | 4.30 | 0.175 |
| | | | Other (S=3) | **-0.54** | 1.25 | 0.068 | **-0.86** | 2.06 | 0.113 |
| | | | Own (S=9) | **-0.72** | 2.49 | 0.049 | **-1.12** | 3.95 | 0.077 |
| | | | Total (S=39) | **-0.52** | 2.37 | 0.023 | **-0.82** | 3.89 | 0.038 |

*Note*. Boldface coefficients are significant at $p < .05$. Annualized outcomes control for number of days between assessments, dDays, as: $dx/Yr = dx \cdot 365/dDays$.

Table 2. *Summary Statistics for Number of Days Between Assessments, dDays.*

| | Subsample | *N* | Mean | Standard Deviation | Std. Error | Quartile values: 25th | Quartile values: Median | Quartile values: 75th | Minimum | Maximum |
|---|---|---|---|---|---|---|---|---|---|---|
| | Control (S=9) | 3,513 | 206.8 | 26.1 | 0.44 | 182 | 195 | 233 | 146 | 245 |
| | ASP (S=30) | 6,693 | 235.7 | 18.4 | 0.22 | 225 | 233 | 245 | 143 | 283 |
| ASP Subsample | HOPS (S=9) | 2,066 | 232.2 | 22.5 | 0.49 | 223 | 224 | 237 | 176 | 279 |
| ASP Subsample | Spark (S=7) | 1,069 | 228.8 | 13.2 | 0.41 | 225 | 231 | 234 | 181 | 249 |
| ASP Subsample | Catch (S=2) | 601 | 265.0 | 2.4 | 0.10 | 263 | 263 | 267 | 260 | 283 |
| ASP Subsample | Other (S=3) | 331 | 229.5 | 9.4 | 0.52 | 228 | 230 | 230 | 182 | 257 |
| ASP Subsample | Own (S=9) | 2,626 | 235.4 | 12.3 | 0.24 | 230 | 234 | 243 | 143 | 271 |
| | Total (S=39) | 10,206 | 225.8 | 25.4 | 0.25 | 215 | 230 | 238 | 143 | 283 |

| Difference between means | Control | HOPS | Spark | Catch | Other | Own |
|---|---|---|---|---|---|---|
| Control | - | < .001 | < .001 | < .001 | < .001 | < .001 |
| HOPS | -25.3 | - | < .001 | < .001 | .027 | < .001 |
| Spark | -21.9 | 3.4 | - | < .001 | .552 | < .001 |
| Catch | -58.1 | -32.8 | -36.1 | - | < .001 | < .001 |
| Other | -22.6 | 2.6 | -0.7 | 35.4 | - | < .001 |
| Own | -28.5 | -3.3 | -6.6 | 29.6 | -5.9 | - |

*Note.* Mean difference below diagonal, difference between mean *p* values above.

Table 3. *Difference between Means Tests: Control versus Various Programs for Five Outcome Measures.*

| Program | Type | Nominal Outcome Differences | | | | | Annualized Outcome Differences | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Outcome Measure | Mean Difference | Significance Level | 95% CI Lower | 95% CI Upper | Outcome Measure | Mean Difference | Significance Level | 95% CI Lower | 95% CI Upper |
| ASP (S=30) | Health Outcome Measures | ΔBody mass index (dB) | 0.08 | .009 | 0.02 | 0.14 | ΔB per year (dB/Yr) | 0.23 | < .001 | 0.12 | 0.33 |
| HOPS (S=9) | | | 0.21 | < .001 | 0.12 | 0.29 | | 0.42 | < .001 | 0.29 | 0.56 |
| Spark (S=7) | | | 0.28 | < .001 | 0.17 | 0.38 | | 0.52 | < .001 | 0.36 | 0.69 |
| Catch (S=2) | | | 0.18 | .006 | 0.05 | 0.32 | | 0.42 | < .001 | 0.21 | 0.63 |
| Other (S=3) | | | 0.00 | .997 | -0.17 | 0.17 | | 0.08 | .585 | -0.20 | 0.35 |
| Own (S=9) | | | -0.11 | .005 | -0.19 | -0.03 | | -0.07 | .256 | -0.20 | 0.05 |
| ASP (S=30) | | ΔB percentile (dB%) | 0.42 | .066 | -0.03 | 0.87 | ΔB% per year (dB%/Yr) | 0.70 | .065 | -0.04 | 1.45 |
| HOPS (S=9) | | | 1.08 | < .001 | 0.47 | 1.68 | | 1.81 | < .001 | 0.84 | 2.79 |
| Spark (S=7) | | | 1.99 | < .001 | 1.22 | 2.75 | | 3.20 | < .001 | 1.98 | 4.43 |
| Catch (S=2) | | | 1.02 | .039 | 0.05 | 1.98 | | 1.36 | .084 | -0.18 | 2.91 |
| Other (S=3) | | | -0.82 | .203 | -2.07 | 0.44 | | -1.33 | .196 | -3.35 | 0.69 |
| Own (S=9) | | | -0.72 | .013 | -1.28 | -0.15 | | -1.08 | .019 | -1.99 | -0.18 |
| ASP (S=30) | Behaviorial Outcome Measures | ΔCurl-ups (dC) | -5.08 | < .001 | -5.62 | -4.55 | ΔC per year (dC/Yr) | -7.78 | < .001 | -8.69 | -6.87 |
| HOPS (S=9) | | | -6.91 | < .001 | -7.66 | -6.17 | | -10.81 | < .001 | -12.01 | -9.61 |
| Spark (S=7) | | | -6.90 | < .001 | -7.84 | -5.96 | | -10.75 | < .001 | -12.26 | -9.23 |
| Catch (S=2) | | | -7.93 | < .001 | -9.12 | -6.75 | | -10.74 | < .001 | -12.66 | -8.83 |
| Other (S=3) | | | -4.80 | < .001 | -6.35 | -3.26 | | -7.61 | < .001 | -10.10 | -5.12 |
| Own (S=9) | | | -2.29 | < .001 | -2.98 | -1.60 | | -3.52 | < .001 | -4.64 | -2.40 |
| ASP (S=30) | | ΔPush-ups (dP) | -1.05 | < .001 | -1.42 | -0.67 | ΔP per year (dP/Yr) | -1.02 | .002 | -1.66 | -0.38 |
| HOPS (S=9) | | | -1.76 | < .001 | -2.25 | -1.27 | | -2.17 | < .001 | -2.97 | -1.37 |
| Spark (S=7) | | | -0.07 | .812 | -0.69 | 0.54 | | 0.38 | .458 | -0.63 | 1.39 |
| Catch (S=2) | | | 1.88 | < .001 | 1.10 | 2.66 | | 3.68 | < .001 | 2.40 | 4.95 |
| Other (S=3) | | | -0.20 | .705 | -1.21 | 0.82 | | 0.29 | .735 | -1.37 | 1.94 |
| Own (S=9) | | | -1.66 | < .001 | -2.11 | -1.21 | | -1.92 | < .001 | -2.66 | -1.18 |
| ASP (S=30) | | ΔMile time (dM) | 0.78 | < .001 | 0.69 | 0.87 | ΔM per year (dM/Yr) | 1.23 | < .001 | 1.07 | 1.39 |
| HOPS (S=9) | | | 1.16 | < .001 | 1.03 | 1.29 | | 1.83 | < .001 | 1.63 | 2.04 |
| Spark (S=7) | | | 0.88 | < .001 | 0.72 | 1.04 | | 1.43 | < .001 | 1.16 | 1.69 |
| Catch (S=2) | | | -0.29 | .004 | -0.49 | -0.09 | | -0.41 | .015 | -0.74 | -0.08 |
| Other (S=3) | | | 0.53 | < .001 | 0.27 | 0.79 | | 0.85 | < .001 | 0.42 | 1.28 |
| Own (S=9) | | | 0.71 | < .001 | 0.59 | 0.83 | | 1.11 | < .001 | 0.92 | 1.30 |

*Note* . CI = confidence interval. Mean difference is Control mean - Program mean using nominal and annualized means from Table 1.

Table 4. *Disaggregating Sandardized Mean Difference (SMD) of Annualized Outcomes by Sex*

| Program Subsample *N* | Outcome Metric, x | Female Ave.dx/Yr | Female SMDx | Female LB | Female UB | Male Ave.dx/Yr | Male SMDx | Male LB | Male UB |
|---|---|---|---|---|---|---|---|---|---|
| HOPS (S=9) | dB/Yr | 0.26 | **0.20** | 0.12 | 0.28 | 0.25 | **0.14** | 0.07 | 0.22 |
| Female | dB%/Yr | -2.40 | **0.19** | 0.11 | 0.26 | -2.34 | 0.02 | -0.05 | 0.10 |
| 987 | dC/Yr | 11.1 | **0.47** | 0.39 | 0.54 | 14.5 | **0.49** | 0.41 | 0.56 |
| Male | dP/Yr | 4.83 | **0.08** | 0.01 | 0.16 | 7.97 | **0.19** | 0.12 | 0.27 |
| 1,079 | dM/Yr | -1.82 | **0.45** | 0.37 | 0.52 | -1.87 | **0.50** | 0.42 | 0.57 |
| Spark (S=7) | dB/Yr | 0.31 | **0.18** | 0.08 | 0.27 | -0.01 | **0.25** | 0.15 | 0.35 |
| Female | dB%/Yr | -2.07 | **0.17** | 0.07 | 0.26 | -5.56 | **0.20** | 0.10 | 0.29 |
| 551 | dC/Yr | 9.57 | **0.40** | 0.30 | 0.49 | 16.2 | **0.56** | 0.46 | 0.65 |
| Male | dP/Yr | 1.58 | **-0.16** | -0.26 | -0.07 | 6.41 | 0.10 | 0.00 | 0.19 |
| 518 | dM/Yr | -1.39 | **0.34** | 0.25 | 0.43 | -1.49 | **0.40** | 0.30 | 0.49 |
| Catch (S=2) | dB/Yr | 0.30 | **0.18** | 0.06 | 0.30 | 0.22 | **0.16** | 0.04 | 0.28 |
| Female | dB%/Yr | -2.17 | **0.17** | 0.05 | 0.29 | -1.67 | -0.01 | -0.14 | 0.11 |
| 304 | dC/Yr | 9.08 | **0.37** | 0.25 | 0.49 | 16.6 | **0.57** | 0.45 | 0.69 |
| Male | dP/Yr | -0.56 | **-0.32** | -0.44 | -0.20 | 1.84 | **-0.19** | -0.31 | -0.07 |
| 297 | dM/Yr | 0.98 | **-0.25** | -0.37 | -0.13 | -0.20 | 0.05 | -0.07 | 0.17 |
| Other (S=3) | dB/Yr | 0.59 | 0.06 | -0.10 | 0.22 | 0.61 | -0.002 | -0.16 | 0.16 |
| Female | dB%/Yr | 0.77 | 0.001 | -0.16 | 0.16 | 0.77 | -0.15 | -0.30 | 0.01 |
| 162 | dC/Yr | 7.30 | **0.29** | 0.13 | 0.44 | 11.91 | **0.38** | 0.23 | 0.53 |
| Male | dP/Yr | 3.80 | 0.005 | -0.15 | 0.16 | 4.22 | -0.04 | -0.20 | 0.12 |
| 169 | dM/Yr | -1.03 | **0.25** | 0.09 | 0.41 | -0.70 | **0.19** | 0.03 | 0.34 |
| Own (S=9) | dB/Yr | 0.88 | -0.05 | -0.12 | 0.02 | 0.62 | -0.01 | -0.08 | 0.06 |
| Female | dB%/Yr | 1.88 | -0.06 | -0.13 | 0.01 | -0.79 | -0.06 | -0.13 | 0.01 |
| 1,292 | dC/Yr | 6.46 | **0.24** | 0.17 | 0.31 | 4.71 | **0.08** | 0.01 | 0.15 |
| Male | dP/Yr | 5.59 | **0.14** | 0.07 | 0.21 | 6.83 | **0.12** | 0.05 | 0.19 |
| 1,334 | dM/Yr | -0.97 | **0.24** | 0.17 | 0.31 | -1.27 | **0.34** | 0.27 | 0.41 |

| Aggregated measures | | Female Ave. SMD | Female *N* sig. Outcomes Pos. | Female *N* sig. Outcomes Neg. | Male Ave. SMD | Male *N* sig. Outcomes Pos. | Male *N* sig. Outcomes Neg. |
|---|---|---|---|---|---|---|---|
| By Program | | Ave. SMD | Pos. | Neg. | Ave. SMD | Pos. | Neg. |
| Across Outcomes | HOPS (S=9) | 0.28 | 5 | 0 | 0.27 | 4 | 0 |
| Across Outcomes | Spark (S=7) | 0.18 | 4 | 1 | 0.30 | 4 | 0 |
| Across Outcomes | Catch (S=2) | 0.03 | 3 | 2 | 0.12 | 2 | 1 |
| Across Outcomes | Other (S=3) | 0.12 | 2 | 0 | 0.08 | 2 | 0 |
| Across Outcomes | Own (S=9) | 0.10 | 3 | 0 | 0.09 | 3 | 0 |
| By Outcome x | | Wtd. Ave. SMDx | | | Wtd. Ave. SMDx | | |
| Across ASP Programs | dB/Yr | 0.09 | 3 | 0 | 0.09 | 3 | 0 |
| Across ASP Programs | dB%/Yr | 0.07 | 3 | 0 | 0.00 | 1 | 0 |
| Across ASP Programs | dC/Yr | 0.35 | 5 | 0 | 0.34 | 5 | 0 |
| Across ASP Programs | dP/Yr | 0.02 | 2 | 2 | 0.11 | 2 | 1 |
| Across ASP Programs | dM/Yr | 0.27 | 4 | 1 | 0.36 | 4 | 0 |
| Total SMD by Gender | | 0.16 | 17 | 3 | 0.18 | 15 | 1 |

| Annualized average outcome changes at control schools by gender, dx/Yr | | Control *N* | dB/Yr | dB%/Yr | dC/Yr | dP/Yr | dM/Yr |
|---|---|---|---|---|---|---|---|
| | Female | 1,769 | 0.75 | 0.79 | 1.44 | 3.73 | -0.03 |
| | Male | 1,744 | 0.60 | -1.93 | 2.67 | 4.88 | 0.01 |

*Note*. dx/Yr = change in x per year. B = body mass index. B% = B percentile. C = curl-up. M = mile. P = push-up. sig. = significant. Ave. = average. Wtd. = weighted (by *N*). LB = 95% SMD lower bound. UB = 95% SMD upper bound. Boldface SMDs are signficant at the .05 level. An SMDx > 0 signifies higher performance by program schools than control schools on x.

Table 5. *Regression Analysis of Annualized Change in Body Mass Index Percentile (dB%/Yr) for At-Risk Students by Program*

| Subsample (*N*) | ASP+ (4,187) | | HOPS+ (2,209) | | Spark+ (1,863) | | Catch+ (1,710) | | Other+ (1,569) | | Own+ (2,556) | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Variable Model: | I | II | I | II | I | II | I | II | I | II | I | II |
| Intercept | **-43.0** | **-34.7** | **-51.4** | -44.7 | -33.3 | -28.6 | -43.7 | -39.8 | -19.1 | -15.0 | -10.9 | -1.9 |
| | (16.0) | (15.9) | (23.1) | (22.8) | (23.7) | (23.5) | (23.6) | (23.4) | (24.9) | (24.7) | (19.5) | (19.3) |
| B% Fall | **9.19** | **8.74** | **9.53** | **9.16** | **9.26** | **9.01** | **9.75** | **9.55** | **8.50** | **8.30** | **7.84** | **7.37** |
| | (0.74) | (0.73) | (1.06) | (1.05) | (1.12) | (1.11) | (1.13) | (1.13) | (1.17) | (1.16) | (0.90) | (0.89) |
| B% Fall$^2$/10 | **-1.85** | **-1.78** | **-1.89** | **-1.84** | **-1.89** | **-1.85** | **-1.96** | **-1.93** | **-1.77** | **-1.74** | **-1.66** | **-1.58** |
| | (0.11) | (0.11) | (0.15) | (0.15) | (0.16) | (0.16) | (0.17) | (0.17) | (0.17) | (0.17) | (0.13) | (0.13) |
| B% Fall$^3$/1000 | **0.97** | **0.94** | **0.99** | **0.97** | **1.00** | **0.98** | **1.03** | **1.01** | **0.94** | **0.93** | **0.88** | **0.85** |
| | (0.05) | (0.05) | (0.07) | (0.07) | (0.08) | (0.07) | (0.08) | (0.08) | (0.08) | (0.08) | (0.06) | (0.06) |
| dPF/Yr: dCurl-ups/Yr/10 | | **-0.24** | | -0.19 | | -0.15 | | **-0.18** | | -0.15 | | **-0.21** |
| | | (0.07) | | (0.10) | | (0.11) | | (0.10) | | (0.12) | | (0.10) |
| dPF/Yr: dPush-ups/Yr/10 | | **-0.56** | | **-0.79** | | **-0.46** | | **-0.54** | | **-0.67** | | **-0.65** |
| | | (0.12) | | (0.16) | | (0.17) | | (0.17) | | (0.17) | | (0.15) |
| dPF/Yr: dMile/Yr/10 | | **2.54** | | **2.57** | | **3.21** | | **2.22** | | **2.44** | | **2.06** |
| | | (0.38) | | (0.52) | | (0.54) | | (0.56) | | (0.59) | | (0.48) |
| Male | -0.15 | 0.10 | 0.06 | 0.29 | -0.87 | -0.71 | -0.72 | -0.51 | -0.30 | -0.19 | -0.09 | 0.05 |
| | (0.31) | (0.31) | (0.43) | (0.43) | (0.46) | (0.45) | (0.47) | (0.47) | (0.48) | (0.48) | (0.40) | (0.40) |
| Program | **-1.33** | **-0.74** | **-2.55** | **-1.58** | **-1.29** | -0.61 | **-1.76** | **-1.86** | 0.72 | 1.11 | -0.60 | -0.14 |
| | (0.33) | (0.33) | (0.45) | (0.47) | (0.54) | (0.55) | (0.64) | (0.65) | (0.84) | (0.84) | (0.40) | (0.40) |
| Adjusted $R^2$ | .387 | .399 | .412 | .427 | .401 | .415 | .405 | .415 | .411 | .425 | .388 | .399 |
| *F* | **530** | **349** | **311** | **207** | **250** | **166** | **234** | **152** | **218** | **144** | **323** | **211** |
| Program Net | Female | Male | Female | Male | Female | Male | Female | Male | Female | Male | Female | Male |
| Effect by Sex | -1.24 | -1.32 | -2.44 | -2.35 | -1.13 | -2.20 | -1.48 | -2.60 | 0.60 | 0.59 | -0.60 | -0.54 |

*Note*. B% = body mass index percentile. dx/Yr = change in x per year. PF = physical fitness. The dependent variable is dB%/Yr, raw regression coefficients (with *SE* beneath). Boldface coefficients are significant at $p < .05$. All subsamples restricted to at-risk students (who had B% ≥ 85 in the fall or spring assessment) and all include 1,430 At-risk control students (Program = 0 for these students). Model I controls for gender and B% Fall. Model II includes three dPF/Yr covariates. *F* significant at $p < .001$ for all models. Program Net Effect is calculated using model II evaluated at At-risk subsample mean dPF/Yr program and control values by sex from Table 6.

Table 6. *Standardized Mean Difference (SMD) of Behavioral Outcomes × At-Risk Status × Sex*

| Program Subsample | *n* | At-Risk Status | dPF/Y | Female mean | Female SMD | Female LB | Female UB | Male mean | Male SMD | Male LB | Male UB |
|---|---|---|---|---|---|---|---|---|---|---|---|
| HOPS (S = 9) | | At-risk | dC/Yr | 11.3 | **0.53** | 0.41 | 0.66 | 13.3 | **0.47** | 0.35 | 0.59 |
| Female | 380 | | dP/Yr | 3.98 | **0.13** | 0.01 | 0.26 | 7.73 | **0.25** | 0.12 | 0.37 |
| Male | 399 | | dM/Yr | -1.99 | **0.50** | 0.38 | 0.62 | -2.08 | **0.55** | 0.43 | 0.67 |
| | | Not At-risk | dC/Yr | 10.9 | **0.43** | 0.33 | 0.52 | 15.2 | **0.49** | 0.40 | 0.59 |
| Female | 607 | | dP/Yr | 5.36 | 0.05 | -0.05 | 0.15 | 8.12 | **0.16** | 0.07 | 0.26 |
| Male | 680 | | dM/Yr | -1.70 | **0.41** | 0.31 | 0.51 | -1.75 | **0.46** | 0.36 | 0.55 |
| Spark (S = 7) | | At-risk | dC/Yr | 9.32 | **0.43** | 0.29 | 0.58 | 15.8 | **0.57** | 0.42 | 0.72 |
| Female | 239 | | dP/Yr | 1.26 | -0.10 | -0.25 | 0.04 | 6.14 | 0.13 | -0.02 | 0.29 |
| Male | 194 | | dM/Yr | -1.30 | **0.34** | 0.19 | 0.48 | -1.68 | **0.45** | 0.30 | 0.61 |
| | | Not At-risk | dC/Yr | 9.77 | **0.37** | 0.25 | 0.50 | 16.5 | **0.55** | 0.42 | 0.67 |
| Female | 312 | | dP/Yr | 1.82 | **-0.20** | -0.32 | -0.07 | 6.57 | 0.07 | -0.05 | 0.20 |
| Male | 324 | | dM/Yr | -1.45 | **0.35** | 0.22 | 0.47 | -1.38 | **0.36** | 0.23 | 0.48 |
| Catch (S = 2) | | At-risk | dC/Yr | 7.56 | **0.35** | 0.17 | 0.52 | 17.2 | **0.63** | 0.45 | 0.81 |
| Female | 148 | | dP/Yr | -0.79 | **-0.28** | -0.45 | -0.10 | 3.26 | -0.07 | -0.26 | 0.11 |
| Male | 132 | | dM/Yr | 1.60 | **-0.36** | -0.54 | -0.19 | 0.09 | 0.02 | -0.16 | 0.20 |
| | | Not At-risk | dC/Yr | 10.5 | **0.41** | 0.24 | 0.57 | 16.1 | **0.53** | 0.37 | 0.69 |
| Female | 156 | | dP/Yr | -0.35 | **-0.35** | -0.51 | -0.18 | 0.71 | **-0.27** | -0.43 | -0.10 |
| Male | 165 | | dM/Yr | 0.39 | -0.13 | -0.30 | 0.03 | -0.42 | 0.09 | -0.07 | 0.25 |
| Other (S = 3) | | At-risk | dC/Yr | 9.12 | **0.42** | 0.16 | 0.69 | 11.1 | **0.37** | 0.15 | 0.60 |
| Female | 57 | | dP/Yr | 3.61 | 0.10 | -0.16 | 0.37 | 3.51 | -0.05 | -0.28 | 0.17 |
| Male | 82 | | dM/Yr | -1.12 | **0.29** | 0.03 | 0.56 | -0.82 | **0.24** | 0.02 | 0.47 |
| | | Not At-risk | dC/Yr | 6.32 | **0.21** | 0.01 | 0.41 | 12.7 | **0.39** | 0.18 | 0.61 |
| Female | 105 | | dP/Yr | 3.90 | -0.05 | -0.25 | 0.15 | 4.90 | -0.02 | -0.24 | 0.19 |
| Male | 87 | | dM/Yr | -0.98 | **0.22** | 0.03 | 0.42 | -0.57 | 0.13 | -0.08 | 0.35 |
| Own (S = 9) | | At-risk | dC/Yr | 5.33 | **0.23** | 0.12 | 0.34 | 5.08 | **0.12** | 0.01 | 0.22 |
| Female | 532 | | dP/Yr | 5.60 | **0.27** | 0.16 | 0.39 | 6.13 | **0.13** | 0.02 | 0.24 |
| Male | 594 | | dM/Yr | -0.65 | **0.18** | 0.07 | 0.29 | -1.10 | **0.31** | 0.21 | 0.42 |
| | | Not At-risk | dC/Yr | 7.24 | **0.25** | 0.16 | 0.34 | 4.42 | 0.06 | -0.03 | 0.15 |
| Female | 760 | | dP/Yr | 5.58 | 0.07 | -0.03 | 0.16 | 7.39 | **0.12** | 0.03 | 0.22 |
| Male | 740 | | dM/Yr | -1.20 | **0.28** | 0.19 | 0.37 | -1.40 | **0.36** | 0.27 | 0.45 |

| Aggregated SMD measures | | Female: At-risk | Not At-risk | Total | Male: At-risk | Not At-risk | Total |
|---|---|---|---|---|---|---|---|
| By Program: Average Across Three Behavioral Outcomes | HOPS (S = 9) | 0.39 | 0.30 | 0.33 | 0.42 | 0.37 | 0.39 |
| | Spark (S = 7) | 0.22 | 0.17 | 0.19 | 0.39 | 0.32 | 0.35 |
| | Catch (S = 2) | -0.10 | -0.02 | -0.07 | 0.19 | 0.12 | 0.15 |
| | Other (S = 3) | 0.27 | 0.13 | 0.18 | 0.19 | 0.17 | 0.17 |
| | Own (S = 9) | 0.23 | 0.20 | 0.21 | 0.19 | 0.18 | 0.18 |
| Using ASP (S=30) Partitioned by At-risk Status × Sex | dC/Yr | 0.30 | 0.34 | 0.31 | 0.35 | 0.35 | 0.38 |
| | dP/Yr | 0.07 | -0.02 | 0.02 | 0.12 | 0.10 | 0.13 |
| | dM/Yr | 0.28 | 0.29 | 0.26 | 0.41 | 0.34 | 0.33 |
| Average dPF/Yr SMD at ASP schools | | 0.22 | 0.20 | 0.20 | 0.29 | 0.26 | 0.28 |

| Control schools *n* & outcome means by At-risk Status × Sex | | Female *n* | dC/Yr | dP/Yr | dM/Yr | Male *n* | dC/Yr | dP/Yr | dM/Yr |
|---|---|---|---|---|---|---|---|---|---|
| | At-risk: | 723 | 0.67 | 2.42 | 0.10 | 707 | 2.37 | 4.26 | 0.18 |
| | Not At-risk: | 1,046 | 1.98 | 4.63 | -0.12 | 1,037 | 2.86 | 5.30 | -0.10 |

*Note.* dx/Yr = change in x per year. C = curl-ups. P = push-ups. M = mile. S = schools. LB = 95% SMD lower bound. UB = 95% SMD upper bound. At-risk = body mass index percentile ≥ 85 in fall or spring. Boldface SMDs are signficantly different from control at $p < .05$. An SMDx > 0 signifies higher x performance at program than control schools.

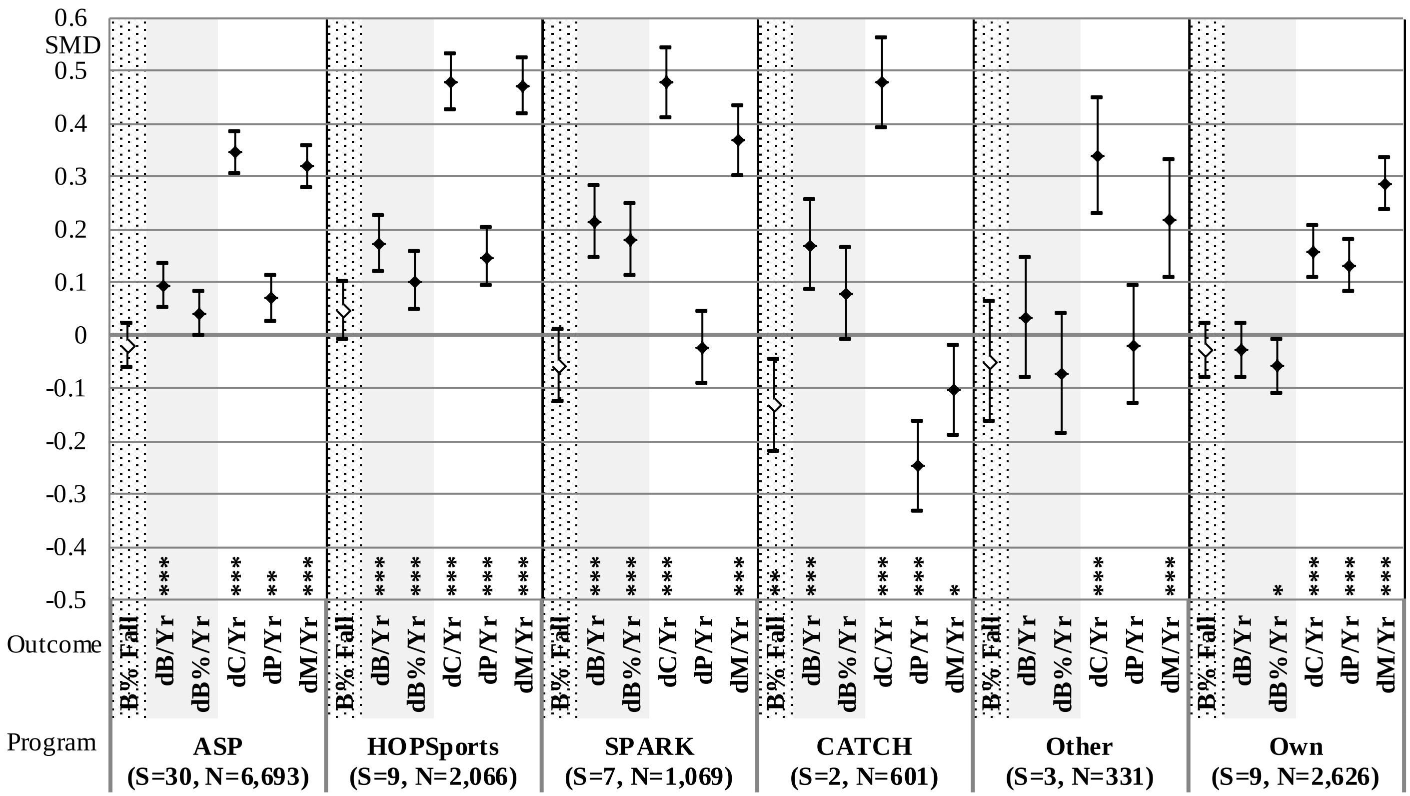


*Figure 1*. Effect size and 95% CI for two health and three behavioral outcomes by program. Effect size measured with standardized mean difference (SMD) calculated as SMD = (mean difference)/(total standard deviation) using annualized outcome mean differences from Table 3 and total standard deviations from Table 1. dX/Yr = change in X per year. B = body mass index. B% = B percentile. C = curl-ups. M = Mile. P = push-ups. Differences taken so that an SMD > 0 means higher performance by program than control students ($N$ =3,513) for that variable (Program - Control for C and P, all else [including B%Fall] is Control - Program). The two health outcomes are shaded gray and the Fall BMI% SMD is speckled. Significant $p$ values noted with * $p$ < .05, ** $p$ < .01, *** $p$ < .001 next to label.

# REFERENCES

Balistreri, K., & Hook, J. (2011). Trajectories of overweight among US school children: A focus on social and economic characteristics. *Maternal & Child Health Journal*, *15*(5), 610-619. doi:10.1007/s10995-010-0622-7

Centers for Disease Control and Prevention. (2009). Children's BMI tool for schools. Retrieved April 14, 2010, 2010, from http://www.cdc.gov/healthyweight/assessing/bmi/childrens_bmi/tool_for_schools.html

Department of Health, Pennsylvania. (2011). Growth screens & BMI-for-age percentile. Retrieved December 31, 2013, from http://www.portal.state.pa.us/portal/server.pt?open=514&objID=556724&mode=2

Department of Health, Pennsylvania. (2013). Procedures for the growth screening program for PA's school-age population manual. Retrieved December 31, 2013, from http://www.portal.state.pa.us/portal/server.pt/community/schools/14130/mandated_school_health_program_(exams_screens)/556692

Du Toit, D., Pienaar, A. E., & Truter, L. (2011). Relationship between physical fitness and academic performance in South African children. *South African Journal for Research in Sport, Physical Education & Recreation (SAJR SPER), 33*(3), 23-35.

Ennis, C. D. (2006). Curriculum: Forming and reshaping the vision of physical education in a high need, low demand world of schools. *Quest (00336297), 58*(1), 41.

Florin, T., A., Shults, J., & Stettler, N. (2011). Perception of overweight is associated with poor academic performance in US adolescents. *Journal of School Health, 81*(11), 663-670. doi:10.1111/j.1746-1561.2011.00642.x

Han, E., & Powell, L. M. (2013). Consumption patterns of sugar-sweetened beverages in the United States. *Journal of the Academy of Nutrition and Dietetics*, *113*(1), 43-53. doi:10.1016/j.jand.2012.09.016

Harris, K. C., Kuramoto, L. K., Schulzer, M., & Retallack, J. E. (2009). Effect of school-based physical activity interventions on body mass index in children: A meta-analysis. *CMAJ : Canadian Medical Association Journal = Journal De l'Association Medicale Canadienne, 180*(7), 719-726. doi:10.1503/cmaj.080966

Heidorn, B. (2011). *National association for sport and physical education, Physical education is critical to educating the whole child [position statement]* Reston, VA.

Katz, D. L. (2009). School-based interventions for health promotion and weight control: Not just waiting on the world to change. *Annual Review of Public Health, 30,* 253-272. doi:10.1146/annurev.publhealth.031308.100307

Katz, D. L., O'Connell, M., Njike, V. Y., Yeh, M., & Nawaz, H. (2008). Strategies for the prevention and control of obesity in the school setting: Systematic review and meta-analysis. *International Journal of Obesity, 32*(12), 1780-1789. doi:10.1038/ijo.2008.158

Morrill, A. C., & Chinn, C. D. (2004). The obesity epidemic in the united states. *Journal of Public Health Policy, 25*(3), 353-366.

Ogden, C. L., & Flegal, K. M. (2010). Changes in terminology for childhood overweight and obesity. *National Health Statistics Reports, (25)*(25), 1-5.

Pérez-Escamilla, R., Obbagy, J. E., Altman, J. M., Essery, E. V., McGrane, M. M., Wong, Y. P., . . . Williams, C. L. (2012). Dietary energy density and body weight in adults and children: A systematic review. *Journal of the Academy of Nutrition and Dietetics, 112*(5), 671-684. doi:10.1016/j.jand.2012.01.020

Piernas, C., & Popkin, B. M. (2011). Increased portion sizes from energy-dense foods affect total energy intake at eating occasions in US children and adolescents: Patterns and trends by age group and sociodemographic characteristics, 1977-2006. *The American Journal of Clinical Nutrition, 94*(5), 1324-1332. doi:10.3945/ajcn.110.008466

Public Law 107-110. (2002). No child left behind act of 2001, H.R. 1. Retrieved from http://www2.ed.gov/policy/elsec/leg/esea02/index.html

Roberts, C. K., Freed, B., & McCarthy, W. J. (2010). Low aerobic fitness and obesity are associated with lower standardized test scores in children. *The Journal of Pediatrics, 156*(5), 711-718, 718.e1. doi:10.1016/j.jpeds.2009.11.039

Shaya, F. T., Flores, D., Gbarayor, C. M., & Wang, J. (2008). School-based obesity interventions: A literature review. *The Journal of School Health, 78*(4), 189-196. doi:10.1111/j.1746-1561.2008.00285.x

Shephard, R. J. (1997). Curricular physical activity and academic performance. *Pediatric Exercise Science, 9*(2), 113-126.

State Board of Education, Pennsylvania. (2010) Annex A. Chapter 12. students and student services: (May 2010).

Telford, R., D., Cunningham, R., B., Fitzgerald, R., Olive, L., S., Prosser, L., Jiang, X., & Telford, R., M. (2012). Physical education, obesity, and academic achievement: A 2-year longitudinal investigation of Australian elementary school children. *American Journal of Public Health*, *102*(2), 368-374. doi:10.2105/AJPH.2011.300220

Torsella, J. (2010). Letter to stakeholders regarding chapter 12. State Board of Education, Pennsylvania. Published November 2, 2010. Retrieved December 31, 2013 from http://www.education.state.pa.us/portal/server.pt/community/current_initiatives/19720/student_health_and_wellness/792455