THE RELATIONSHIP BETWEEN PHYSICAL ACTIVITY AND ACADEMIC PERFORMANCE

by

Corey A. Nunlist

A Project Submitted to the Faculty of

The Evergreen State College

In Partial Fulfillment of the Requirements

for the Degree

Master in Teaching

2013

This Project for the Master in Teaching Degree

by

Corey A. Nunlist

has been approved for

The Evergreen State College

by

Jon Davies, Member of the Faculty

June 2013

ACKNOWLEDGEMENTS

Thank you everybody and everything, especially all the good bodies and good things, and especially especially my parents.

ABSTRACT

Schools are increasingly scaling back physical education programs in favor of increasing time spent on subjects such as reading and math. However, the link between physical activity and academic performance is not well understood. This review seeks to inform the discussion on that link to assist teachers and policymakers in making decisions regarding the school day. Results suggest physical activity, including school physical education programs, has no effect on academic performance and that certain types of coordinated activity may improve student concentration and attention. Implications for teaching include taking short breaks during long subject lessons and capitalizing on student attention immediately after breaks. More rigorous and varied research is needed, including objective and qualitative studies.

TABLE OF CONTENTS

CHAPTER 1: INTRODUCTION

Introduction

The physical health of children in the United States is today worse than ever. Childhood obesity more than tripled from 1980 to 2008, and children are consequently suffering from related ailments like diabetes, sleep apnea, and low self-esteem (CDC 2011). Schools provide a unique opportunity to moderate and perhaps reverse this trend: children spend a large percentage of their lives in schools, are constantly around peers with whom they can interact and play, and are around numerous adults who can teach and model good exercise habits.

However, while childhood obesity has risen, time spent on school-based physical activity has fallen (Coe, Pivarnik, Womack, Reeves, & Malina, 2006). Schools have experienced a decrease in funding and are under increasing pressure to raise scores on standardized tests (Ravitch, 2010). The physical benefits of exercise have always been known. More recently, research has established other positive results of physical exercise, such as increased self-esteem and reduced stress and anxiety (U.S. Department of Health and Human Services, 2008). Perhaps not surprisingly, given these results, exercise has also been shown to affect the brain's physiology in positive ways (CDC, 2010).

The link between physical activity and academic performance, however, remains less established. The relationship has recently been studied more but there is still a limited amount of research on the topic and studies' findings have been mixed; some have found a positive effect of physical activity on academic performance (e.g., Shepard, 1996) and some have found a negative effect (e.g., Tremblay, Inman, & Willms, 2000).

This review seeks to inform the discussion on the relationship between physical activity and academic performance through a critical review of the related literature. It will seek to address whether fitness and physical activity are related to or affect academic performance and, if so, to what degree.

Rationale

The debate over time spent on physical activity during school hours has shifted from children's health to children's academic achievement. Despite serious health concerns regarding youth in the United States, policy makers have increasingly focused on school as a place for academic learning in certain core subjects (Ravitch, 2010). Educators and administrators may justify the use of school time for physical activity by providing evidence that such activity improves children's performance in those core subjects. Likewise, if physical activity has a negative effect on students' competence in those areas, the use of school time for such activities must be questioned.

Teachers are ultimately responsible for helping students learn as much of established school curricula as possible, and it is important that teachers know how best to accomplish this. However, teachers have a flexibility and an obligation to address other student needs while maintaining focus on the overarching role of increasing student understanding of curricula; addressing health needs such as diet and nutrition has shown to improve student academic performance and behavior (Storey, Pearce, Ashfield-Watt, Wood, Baines, &

Melson, 2011; Vinciullo & Bradley, 2009). That flexibility and obligation allow teachers to create more active lessons to benefit students if benefits are demonstrated. If active lessons or time spent on physical activity do not affect student performance, then teachers will have to use their own judgment on whether physical activity is worthwhile; given the enormous costs of obesity to both individuals and society (Finkelstein, Trogdon, Cohen, & Dietz, 2009), teachers may be justified in providing physical activity during the school day. Whatever their decision, teachers should justify their actions based on available research and the academic needs of their students.

Teachers are partially limited by the constraints placed upon them by administrators' decisions. Many schools, if given the option, will continue to restrict opportunities for structured or unstructured physical activity in favor of increased instructional time. However, if physical activity has a positive effect on academic performance, then educators should find ways to incorporate activity and exercise into academic lessons. The opportunities to do so are numerous and varied.

Historical Background

Physical education (PE) was largely absent in American schools until the late 19th century, when school-based gymnastics and aerobics programs first appeared. These programs became more prevalent in the United States as the country became more industrialized and, consequently, more sedentary. However, it was not until data was released regarding the poor fitness levels of World War I draftees that PE in schools was given serious attention. This focus was short-lived as the Depression of the 1930s shifted money and attention elsewhere, but again gained steam after World War II draftee fitness reports were examined and were similarly poor. Physical fitness was increasingly studied, resulting in both valid measures of and standards for fitness. Additionally, the Cold War provided the impetus for constant examination of Americans' health and ability to immediately coalesce into a fit army. Numerous developments in the field of physical fitness came about in the 1950s and 60s, including the formation of the President's Council on Youth Fitness and a shift of attention from intervention to prevention (Dalleck & Kravitz, 2002), and school PE programs expanded through the 70s. This trend was reversed in the mid-90s, with the proportion of students attending daily PE classes declining significantly in the first half of the decade, and then remaining at this lower level through 2003 (Centers for Disease Control and Prevention, 2004).

More recognition has been paid to children's physical health in the first part of the 21st century as child obesity rates and associated health problems have risen dramatically (CDC, 2011). However, recent decreases in school funding combined with an increased emphasis on standardized test results has led to the reduction or elimination of numerous opportunities for children to get regular exercise, such as PE classes and recess (Dills, Morgan, & Rotthoff, 2011). Proponents of school-based PE have needed to justify exercise in schools with more than the argument that it improves physical health. Thus, more research has been conducted recently examining the link between physical activity and academic performance, seeking a rationale for providing or not providing opportunities for physical activity during school.

Schools today continue to be scrutinized and evaluated largely based on standardized test results. The validity of this evaluation method is often questioned, with some arguing that such tests improve performance (Phelps, 2011) and some arguing against their use (Popham, 1999). Regardless, these tests play a large role in measuring student, teacher, and school progress, and the resulting pressure ensures decisions on school schedules and curricula will incorporate the expected results on standardized tests. This political pressure often conflicts with societal pressure to improve health and fitness, especially among children. The problem of childhood obesity is widely recognized, as is the power of schools in shaping the habits of children; but policymakers, administrators, teachers, and society at large hesitate to devote school time to health and fitness in the face of immense pressure to increase student academic performance. Today, these actors must decide what they value and how best to mirror those values in the public education system.

Definitions

There are several categories of factors that fall under academic performance in the literature. Broadly, these factors deal not only with traditional indicators of achievement such as grades and test scores, but also things like behavior, motivation, and attention. The Centers for Disease Control and Prevention (CDC) divided academic performance into three categories: cognitive skills and attitudes, academic behaviors, and academic achievement (CDC, 2010, p. 10). This review uses the CDC's definition of academic performance inclusive of these three categories, as well as definitions for PE, physical activity, and recess as follows:

> Cognitive skills and attitudes: Cognitive skills and attitudes include both basic cognitive abilities, such as executive functioning, attention, memory, verbal comprehension, and information processing, as well as attitudes and beliefs that influence academic performance, such as motivation, self concept, satisfaction, and school connectedness.
>
> Academic behaviors: Academic behaviors include a range of behaviors that may have an impact on students' academic performance. Common indicators include on-task behavior, organization, planning, attendance, scheduling, and impulse control.
>
> Academic achievement: Academic achievement includes standardized test scores in subject areas such as reading, math, and language arts; GPAs; classroom test scores; and other formal assessments.
>
> Physical activity: Any bodily movement produced by the contraction of skeletal muscle that increases energy expenditure above a basal level.
>
> Physical education (PE): A curricular area offered in K-12 schools that provides students with instruction on physical activity, health-related fitness, physical competence, and cognitive understanding about physical activity, thereby enabling students to adopt healthy and physically active lifestyles.

Recess: A time during the school day that provides children with the opportunity for active, unstructured or structured, free play.

This definition of recess as a structured or unstructured opportunity for free play contrasts with the idea of exercise, which is defined in a U.S. Department of Health and Human Services Report as, “A subcategory of physical activity that is planned, structured, and repetitive and purposive in the sense that the improvement or maintenance of one or more components of physical fitness is the objective” (U.S. Department of Health and Human Services, 2008, pp. C1-C2). School-based physical activity is that which takes place during the regular hours of school; this excludes extra-curricular activities such as sports and clubs.

Limitations

This review informs the discussion on the relationship between fitness, physical activity, and academic performance of students, and is limited to critically examining the methodologies and results of 30 studies on that relationship. The body of research on this topic consists of more than 30 studies; therefore, research has necessarily been excluded from the review. However, the review includes the highest quality research relating to the research question. The review is anchored in the professional literature; only studies from peer-reviewed journals are included. The review also includes studies the results of which span the range of reported results regarding this topic. Therefore, the review, though limited in breadth, presents a high quality and objective view regarding the relationship between physical activity and academic performance.

Another major limitation to the scope of this review is that it only examines the literature on K-12 students. Additionally, although there are numerous undisputed benefits from physical activity, this review is concerned with the effect of fitness and physical activity on academic performance alone; therefore, improvements in physical health and other non-academic indicators are not considered, regardless of indirect effects that improvement in these areas has on academic performance.

Summary

This chapter has given a brief overview of the debate regarding time allotted to physical activity during regular school hours; the history of school-based physical activity in schools and its current state today; and the issue's importance to administrators, educators, and children. Historically, the amount of time schools allocate to PE has fluctuated, responding to periods of increased concern about the fitness of youth and periods where the focus has been elsewhere, such as on academics. Those two concerns have recently collided, leading to a debate about the role of schools in providing children with opportunities for exercise. Children spend huge amounts of time in school, where they are expected to learn about certain core subjects upon which they will be tested. Some argue that expectations for personal health and fitness reside elsewhere, outside of school hours. There are conflicting views regarding the effect of physical activity on academic performance, and the research on this topic has profound implications for how school time should be allotted. The next chapter will seek to inform this debate by critically reviewing the literature on the relationship between physical activity and academic performance.

CHAPTER 2: CRITICAL REVIEW OF THE LITERATURE

Introduction

Chapter one of this review discussed the role of physical education (PE) in regular classroom settings. Historically, PE was given little attention in the United States until the poor fitness of American youth was exposed during World War I. School was viewed as a medium through which to reverse this trend, and PE gained prominence in American schools through much of the 20th century. Beginning in the 1990s, schools cut back on PE in favor of more instructional time for subjects like reading and math, especially in light of state and federal testing requirements. At the same time, childhood obesity increasingly became an issue of national concern. A discussion of the relationship between physical activity and academic achievement will inform the allocation of valuable school time.

Chapter two critically reviews the literature on the relationship between physical activity and measures of academic performance and is organized into five sections: the relationship between fitness levels and academic performance, the relationship between physical activity and academic achievement, the relationship between exercise and concentration, the effects of fitness interventions on achievement test scores, and the relationship between physical activity and other areas of school performance. Together, these sections review a large portion of the research on the connection between student physical fitness and school performance.

Fitness Levels and Academic Performance

Six studies in this review looked at the connection between students' fitness levels and academic performance. Castelli, Hillman, Buck, and Erwin (2007) studied the relationship between physical fitness and academic achievement of 3rd and 5th graders in a medium-sized urban community and found that certain subscales of a fitness test were positively correlated with academic achievement, while other subscales were not. Chomitz et al. (2009) studied the relationship between the results of Massachusetts standardized achievement tests and fitness tests for 1,841 Cambridge public K-8 students and found that higher fitness was positively and significantly related to better test scores. Grissom (2005) matched fitness test and Stanford Achievement Test data for 884,715 California public school students and found significant positive relationships between the two. Hillman et al. (2005) monitored both high- and low-fit elementary students during a cognitive test and concluded that higher fitness levels were associated with faster response times and more brain activity. Dwyer, Sallis, Blizzard, Lazarus, and Dean (2001) found significant positive associations between academic performance and physical fitness of 2,180 Australian children aged 7-15, but the results varied by subgroups of age and gender. Lastly, Knight and Rizzuto (1993) looked at the relationship between the balance skills and reading and mathematics scores of 122 public school students in Georgia and found that mean achievement scores generally increased as balance skills increased.

Castelli et al. (2007) looked at the relationship between student scores on the Illinois Standards Achievement Test (ISAT) and scores on the Fitnessgram, which measures muscle fitness, aerobic capacity, and body composition. Castelli et al. noted the ISAT has been deemed valid and reliable but did not back up this claim; they did cite research establishing the validity and reliability of the Fitnessgram.

Castelli et al. (2007) conducted their study with 3rd and 5th graders from a medium-sized urban school district in Illinois. Four of the district's 11 schools were chosen: two academically higher-performing schools (76.3% of students met or exceeded math standards and 86.4% met or exceeding reading standards) and two academically low-performing schools (46.2% math and 40.4% reading). Castelli et al. then excluded students with individual education plans that indicated a disability, and any students who did complete the ISAT or all of the Fitnessgram subtests. The final sample (N = 259) had a mean age of 9.5 years, was 51% male, 78% Caucasian, 12% African American, 5% Asian, and 3% Hispanic.

Students completed the ISAT as part of regular school requirements. Researchers administered the Fitnessgram during regular PE classes. Pearson product-moment correlation analyses were conducted on all five Fitnessgram subscales, a composite total fitness score, a composite achievement score, both the reading and math subscales, age, sex, school, and a socioeconomic status (SES) indicator. Castelli et al. (2007) used the associations found to conduct two-step hierarchical regressions that found higher academic achievement scores were positively related to greater total fitness (pr ≥ .43, $t(256) \geq 7.7$, $p < .001$, β ≥ .43). Castelli et al. also found significant associations between higher total academic achievement and lower BMI (pr = -.17, $t(250) = 2.8$, $p < .01$, β = -.16) and better aerobic fitness (pr = .42, $t(250) = 7.3$, $p < .001$, β = .43). No correlations were found between academic achievement and the other three fitness subscales.

Castelli et al. (2007) are transparent and the results have been replicated, most notably in a study by the California Department of Education that also used the Fitnessgram and state achievement test data. Castelli et al. accounted for several variables in their research, such as age, sex, school-wide academic achievement, and SES. However, the Fitnessgram, while more easily generalizable to other public education settings, is not the best measure of physical fitness due to dependence on student motivation to perform. Additionally, time spent in PE was not measured and the assumption that schools in the same district have the same PE practices could easily be flawed; students who performed better on the Fitnessgram may have spent less time being physically active, or vice versa.

Castelli et al. (2007) excluded many students from the final sample because they lacked parental consent, data on academic or fitness tests, or because of disabilities. These factors may have biased the results towards a more fit, higher-achieving sample, threatening the study's internal validity. Another threat to the study's internal validity is that students familiarized themselves with the Fitnessgram a week in advance of testing; this may have resulted in a less accurate measure of fitness.

Chomitz et al. (2009) conducted their study in Cambridge, Massachusetts, public schools, and looked at the relationship between fitness and student performance on the Massachusetts Comprehensive Assessment System (MCAS) math and reading tests; fitness tests were adapted in part from the Fitnessgram. The researchers excluded from their analysis students who did not complete the entire fitness test, students with implausible fitness or academic achievement data, Native American students (due to confidentiality issues associated with a small sample size), and students with identifiers that indicated a high degree of impairment; about 13% of the sample was excluded in all. The final sample consisted of 1,841 4th, 6th, 7th, and 8th grade students; 5th grade students do not take the MCAS reading or math tests. Sixty-five percent of the students were non-white, 45% low-income, and almost 40% overweight or at risk of overweight. Seventy-two percent of the sample achieved a passing score on the math test, 89% passed the English test, and students passed on average 3.6 (out of 5) fitness subtests.

Chomitz et al. (2009) conducted a series of bivariate and multivariate regression analyses on the data and found, after controlling for gender, ethnicity, weight status, grade, and SES, significant and positive relationships between the number of fitness subtests passed and the likelihood of having a passing score on the MCAS math or reading test. The relationship was much stronger for math ($p < .0001$) than for reading ($p < .05$). The researchers estimated that for each additional fitness test passed, the odds of passing the math MCAS increased by 38%, and the odds of passing the reading MCAS increased by 24%.

The internal validity of this study is dampened by the exclusion of several groups of people, including those with disabilities or missing data. Further, the authors noted the internal validity of their study may be compromised due to lack of data controls, specifically with the recording of fitness tests passed. The conclusions reached by Chomitz et al. (2009), including the stronger relationship between math scores and fitness in comparison to reading and fitness, support similar research conducted by Castelli et al. (2007), lending both studies more reliability. The sample's diverse characteristics, representative of many urban school districts in the United States, make findings generalizable to other public education settings. The researchers clearly stated their methods and procedures and cautioned against causal interpretations of their findings, lending the study transparency and objectivity.

Grissom (2005) matched 2002 fitness and achievement test data for 884,715 California public school students. California used the Fitnessgram to annually assess fitness. Achievement was measured by the norm-referenced Stanford Achievement Test 9th Edition (SAT/9) for reading and math. Grissom calculated normal curve equivalent scores to look at the relationship between SAT/9 and Fitnessgram scores by number of fitness subtests passed. He found the number of passed fitness tests increased with the normal curve equivalent scores for both reading and math. Grissom then tested the relationship using analyses of variance and found it to be significant for both reading and math (*p* <

.0001). Grissom looked at the relationship further with Tukey's studentized range test for multiple comparisons and found that the relationship was significant for each number of fitness tests passed compared with the next number. For instance, achievement scores for students who passed all six fitness tests were significantly higher than for students who passed five tests. This was true for both reading and math scores ($p < .01$). Grissom found the same positive significant relationships after controlling for gender and SES.

Grissom's (2005) findings have high external validity given the study's inclusion of all California public school students with valid data, and conclusions can be generalized to other state populations. Missing data may have biased results, but there is no specific reason to believe so. However, as Grissom himself notes, variables not examined in the study could be the cause of the significant relationships found. Grissom is also clear in stating the relationships as correlational and not causal, and that a discussion could easily be framed around the positive relationships found between SES and both academic achievement and physical fitness.

Hillman et al. (2005) studied response times and accuracy of two groups of Illinois children ($N = 600$), one classified as high-fit and the other low-fit, as they completed a cognitive test. Hillman et al. selected the high-fit group from the top 10% of Fitnessgram results and the low-fit group from the bottom 10%. The high fit children averaged 9.3 years of age and an IQ of 116.7 (as measured by the K-BIT), and had 5.0 years of education. Students in the low-fit group averaged 9.8 years of age, an IQ of 108.7, and 5.3 years of education. The groups were equivalent on SES based on participation in the free or reduced-price lunch program, highest level of education obtained by the parents, and the number of parents who worked fulltime.

Students in the study were asked to press a button when a target stimulus was flashed (one trial). Student response time and accuracy were measured over three trials. Hillman et al. (2005) used analyses of variance to look at the interactions among age, fitness, and condition (target or non-target stimulus) for response accuracy, and at the interactions between age and fitness for response time. Because the overall study included two high- and low-fit adult groups, differences between the two groups of children were examined with *post hoc* analyses that used the Bonferroni correction procedure, a more stringent test for significance. Hillman et al. found that high-fit children had faster response times than low-fit children ($t(1,22) = 2.8$, $p = .01$) and that differences on response accuracy were non-significant for both target stimuli ($t(1,49) = 2.1$, $p = .04$) and non-target stimuli ($t(1,49) = 1.5$, $p = .14$).

Hillman et al.'s (2005) study has threats to both its internal and external validity. The study sample was not drawn randomly from the elementary school population; participants were selected solely on the basis of their performance on the Fitnessgram, which, though measured as fairly valid and consistent with more objective measures of fitness, is itself a field test and more subject to observer bias. Hillman et al. mitigated this concern by measuring the groups' age, height, weight, IQ, and SES. The groups were not significantly different on these measures ($p \leq .05$), but there was a difference in IQ that might have been classified as significant in a sample size larger than 24. Additionally, Hillman et al. collected no demographic information for four subjects, or 17% of the sample. Lastly, the authors stated that the groups are equivalent on the SES indicator but did not say how it was measured. However, Hillman et al.'s (2005) objectivity is strengthened by their cautious approach to determining significance, as evidenced by their use of the Bonferroni correction, and because the findings were presented openly. A major part of the study conducted by Hillman et al. was the measurement of brain activity in subjects as they completed the cognitive tests. Unfortunately, a critique of the methods used for this part of the study is outside the scope of this review. It is worth noting, however, that Hillman et al. reported the high-fit children had significantly more brain activity than all of the other groups, indicating the use of more neurons in completing the cognitive tests.

Dwyer et al. (2001) used data from the Australian Schools Health and Fitness Survey (ASHFS) to look at the association between academic performance and physical fitness in Australian children aged 7-15. The ASHFS, conducted in 1985, used a two-step sampling procedure to get a random sample of schools by size and location and a random sample of boys and girls at each age with a goal of 9,000 students in all. Trained personnel visited schools over a three-month period and tested students on a variety of fitness indicators such as lung capacity, flexibility, and running ability. Data was also collected on student demographics, activity in and out of school, self-assessed scholastic ability, and school representative-assessed scholastic ability.

Dwyer et al. (2001) first calculated Spearman rank correlation coefficients for the associations between students' and representatives' perceptions of scholastic ability. Coefficients ranged from .29 to .51. The researchers cited these coefficients as evidence of a significant relationship between the two measures though coefficients in that range signal weak positive relationships at best. Dwyer et al. then calculated Spearman rank correlation coefficients for the relationship between representatives' ratings of scholastic ability and fitness variables from the ASHFS survey. Coefficients, many of which were labeled as significant by the authors, ranged from -.19 to .27. Finally, Dwyer et al. conducted linear regression analyses looking at the relationship between representatives' ratings of scholastic ability and sit-ups that adjusted for type and location of school, SES, age of student, student physical characteristics, and physical habits of the student such as bedtime and tendency to eat breakfast before school; the authors used sit-ups due to the variable's non-linear relationship with scholastic ability. Dwyer et al. found significant positive relationships between scholastic ratings and sit-up ability for both genders and all age groups, with the exceptions of 9-year-old boys and 10- and 11-year old girls.

The significant findings reported by Dwyer et al. (2001) should be viewed with extreme caution. The authors did not report how significance was determined in either the Spearman rank correlation coefficients or the regression analyses. They admit the Spearman coefficients are low but maintain that such small relationships should be significant because the expected effects of physical fitness on academic ability would be small. The authors did not explain the weak connections between students' and representatives' ratings of scholastic ability. The strong voice used to report weak connections calls into question the objectivity of the authors.

There were other significant threats to the internal validity of the study conducted by Dwyer et al. (2001). School representatives cannot reasonably be expected to accurately assess the abilities of all their students, particularly in the case of large schools. Likewise, having students rating themselves as below, at, or above average is an imprecise measure of achievement. The weak connections between the two ratings provide further evidence of the measures' lack of validity. Another threat to the study's internal validity is the use of postal codes as a measure of SES. There was likely a wide variation in students' SES within postal codes, and this is lost when assigning each student the SES of a much larger area. Third, the number of students in the final sample is unclear from both the discussion of the ASHFS survey and the results of the analysis performed by Dwyer et al.

Knight and Rizzuto (1993) studied the relationship between balance skills and reading and math scores on the Iowa Test of Basic Skills for 122 elementary students in the Georgia public school system. The sample included 64 boys and 58 girls from 2nd to 4th grades, aged 7 to 11. The researchers looked at 10 balance skills in all, split into dynamic balance skills that required students to maintain a balanced position while moving, and static balance skills that required students to maintain a stationary position. Balance tests were judged on a pass/fail basis, with each student's score between one and ten depending on how many tests passed. One-way multivariate analysis of variance was used to test the relationship between balance scores and Iowa Test of Basic Skills scores, and results suggested that both math and reading scores increase as balance scores increase.

Knight and Rizzuto (1993) found a positive relationship between better balance and higher achievement but these results should be viewed with extreme caution. The researchers did not report how the balance tests were administered, nor any other information on the study sample or data collection. Because there is such little information given, one cannot hope to replicate this study, and the results have limited external or internal validity.

Five of the six studies in this section found positive relationships between fitness levels and academic performance. Castelli et al. (2007), Chomitz et al. (2009), Dwyer et al. (2001), and Grissom (2005) found that fitness was positively correlated with academic achievement. Knight and Rizzuto (1993) found that student achievement scores generally increased with balance skills. The sixth study, Hillman et al. (2005), found that higher fitness levels were not associated with more accurate responses on a concentration test. These studies were concerned with the relationships between academic performance and fitness. The next section concerns the relationship between physical activity and academic performance as measured by students' concentration and attention.

Physical Activity, Concentration, and Attention

This review looked at four studies regarding the effect of physical activity on tasks of concentration. Pellegrini, Huberty, and Jones (1995) performed three similar experiments involving elementary public school children that studied the connection between recess patterns and subsequent performance on a test of attention and found that students were less attentive before recess and when recess was delayed. Caterino and Polak (1999) administered a concentration test to two groups of children in grades 2-4, with one group taking the test after 15 minutes of stretching and aerobic walking and the other immediately after regular classroom lessons; there were no significant differences in test results between the two conditions for 2nd and 3rd graders, but 4th graders who exercised immediately before the concentration test significantly outperformed their counterparts. Budde, Voelcker-Rehage, Pietrabyk-Kendziorra, Ribeiro, and Tidow (2008) studied the performance of two groups of high school students on a test of attention and concentration and found the group that took the test immediately following 10 minutes of coordinated exercise showed significantly higher improvement on the test compared with the group that engaged in 10 minutes of regular, non-coordinated exercise. Raviv and Low (1990) examined whether the concentration of 11-12 year old students was higher or lower following a PE classes and found that mean concentration scores were elevated towards the end of classes.

Pellegrini et al. (1995) performed three similar experiments to examine how students' activity levels during recess and their amount of sedentary time before recess affect their attention to school tasks after recess. All three experiments took place in the same public elementary school in the southeastern United States. The first experiment included students in kindergarten (11 boys, 6 girls, 5.6 years old), second grade (13 boys, 11 girls, 7.5 years old), and fourth grade (10 boys, 11 girls, 9.7 years old). The authors stated that the subjects were mostly European American and about one third African American, with a few Asian Americans as well, and "represented a variety of social, economic, and cultural backgrounds" (Pellegrini et al., 1995, p. 848), though they present no evidence to support the statement.

In the first experiment, the researchers controlled the timing of recess to extend or shorten the amount of time students sat beforehand. Second and fourth grade students were read a male- or female-preferred (based on the main character) story immediately before and after recess. Kindergarten students did sedentary activities. Attention was measured by student eye contact with the teacher or the activity; Pellegrini et al. (1995) cite research that gaze is a reasonable measure of attention (Pick, Frankel, and Hess, 1975). Physical activity during recess was measured on a nine-point scale that ranged from lying down to running; the authors do not cite information on the validity or reliability of this measure. Five observers, including one reliability judge, coded children's recess behaviors and pre- and post-recess attention over a two-month span using scan sampling and instantaneous recording, with each child observed in each setting (pre-, during-, and post-recess) each day. Observers had about a month to practice the coding procedures, and Cohen's kappa coefficients of > .75 indicated excellent inter-rater reliability for all measures.

Pellegrini et al. (1995) calculated repeated measures analysis of variance on pre-recess attention, analyzing the effects of gender and condition for kindergarten students and adding the effects of grade and gender-preference of task for second and fourth grade students. Kindergarten and fourth grade students were significantly less attentive when the timing of recess was delayed. Pellegrini et al. analyzed post-recess results by correlating recess activity to post-recess attention. The correlation between physical activity at recess and post-recess attention for the entire sample was not significant ($r = .05$, $N = 62$). Pellegrini et al. also analyzed the mean attention scores for pre- and post-recess and found fourth graders ($M_{pre} = 7.45$, $M_{post} = 9.15$) and second graders ($M_{pre} = 6.44$, $M_{post} = 14.83$) were significantly more attentive after the break. However, activity during recess was not significantly connected to post-recess attention scores.

Pellegrini et al. (1995) replicated their first experiment in the same school with second and fourth graders one year later. Similar to experiment one, they found that children's inattention was higher when recess was delayed, and higher pre-recess than at post-recess, though the difference was not significant for fourth graders. They also found, again, that post-recess inattention was unrelated to physical activity. Pellegrini et al. conducted a third experiment with two fourth grade classrooms to test the effects of temperature on recess behavior and attention, with one class having recess inside and the other outside. Results regarding inattention pre- and post-recess and the relation of physical activity to post-recess inattention were similar.

The major strength of the study conducted by Pellegrini et al. (1995) is the replication of their findings from the first experiment in the second, and that they found similar findings after a slightly different third experiment. This gives the study results great reliability. The researchers were also meticulous in their coding of behaviors and attention, training observers beforehand and reporting strong correlations for inter-rater reliability. The external validity of the study is further strengthened because it would be relatively easy to alter the schedule of an elementary classroom in order to give students breaks to increase their attention. The small sample sizes are a threat to the study's internal validity, as is the lack of descriptive statistics and analyses regarding the subjects; the reader does not know much about the population studied other than the grade level and an approximation of race/ethnicity. Nevertheless, the careful methods and reliability of the study provide evidence that children's attention will increase if they are not sedentary for long periods of time.

Caterino and Polak (1999) conducted their study with students in grades two through four. Each grade consisted of two groups: one group that took the Woodcock-Johnson concentration test immediately following regularly scheduled classroom lessons, and another group that immediately took the test following 15 minutes of stretching and aerobic walking. The Woodcock-Johnson test of Concentration requires students to match five symbols in each of 30 rows to a criterion symbol. After using the Scheffé multiple comparison test to establish that the grade four physical activity group performed the best, Caterino and Polak performed a one-way analysis of variance on the fourth graders' mean scores and found the mean scores for the physically active group were significantly higher than those of the non-physically active group (*p* = .05). There were no significant differences in the other grades.

The results of the study by Caterino & Polak (1999) should be viewed with skepticism. The authors do not report any data on the equivalence of the two groups in each grade, nor do they describe what the physical activity group did prior to taking the test other than vaguely saying the students stretched and engaged in aerobic walking. The objectivity of the authors can be questioned when they state in their findings that students in the physical activity group had significantly higher scores than those in the regular classroom setting group; while this is true for the means of the two conditions in aggregate, the researchers themselves noted earlier that the significance was essentially determined by the superior performance of the grade four physically active group. They further state that, "scheduling a PE class need not be of concern up to Grade 4," (Caterino & Polak, 1999, p. 248) though the physical activity condition would hardly be considered a PE class.

Budde et al. (2008) used performance on the d2 Concentration Test (d2) to measure the attention and concentration of a group that took the test immediately after 10 minutes of coordinative exercise compared with a group that took the test after 10 minutes of regular exercise. The d2 requires students to identify target symbols mixed with similar symbols. Budde et al. cite evidence for the good internal reliability, criterion, construct, and predictive validity of the d2. Subjects were 115 students randomly assigned to the two groups, though 16 were later excluded due to highly unusual test results. The 99 remaining subjects were about 15 years old and 80% male. Pretests took place immediately after normal school lessons and posttests took place one week later after the coordinative exercise or normal sport lessons. On test days, students refrained from any exercise prior to testing.

Budde (2008) analyzed results with mixed factor analyses of variance, which revealed a significant group by time interaction. All participants improved from pre- to posttest; the experimental group improved significantly more than the control ($p < .05$). The researchers also collected heart rate information on the two groups and differences were not significant and claimed this as evidence that the coordinative nature of the experimental group exercises was responsible for the significant differences in performance.

Although the study conducted by Budde et al. (2008) randomly assigned subjects to experimental or control conditions, they do not give any information on the randomization process or on group equivalence after randomization; this is a threat to the study's internal validity. Another threat is that the posttest took place a full week after the pretest; although the pretest established a baseline concentration score for the groups, it would have been better performed immediately prior to the exercise. It was not clear why the authors chose their approach. A strength of the study was the reliability and validity of the instruments used, and the carefully controlled nature of the 10 minute exercise sessions; experimental subjects followed a strict program and controls performed exercise of moderate intensity. This was measured and recorded. The results of the study suggested that coordinative exercise can improve concentration, but the external validity of the study is limited by its use of subjects from an elite performance school in Germany. Elite is not defined, but the term suggests a high achieving population with many resources; these are not characteristics of the majority of public school populations.

Raviv and Low (1990) examined the concentration scores of 96 11-12 year old students to determine differences in concentration before and after PE classes, after PE versus after science classes, and before and after morning versus afternoon PE classes. Students were divided into four groups: two science classes and two PE classes. Subjects were taught twice daily by the same teachers at 8 am and 2 pm. Similarly to the study conducted by Budde et al. (2008), the d2 Concentration Test was used to measure concentration.

Raviv and Low (1990) analyzed results with a multiple analysis of variance 2x2x2 factorial design using subject matter, time of day, and time of test. A significant three way interaction was found, with students performing better at the end of early lessons in each subject ($F(1, 92) = 9.17$, $p < .01$). Students in the afternoon PE class showed statistically significant gains from pre- ($M = 323.27$, $SD = 76.18$) to post-class ($M = 397.57$, $SD = 96.60$); gains were not significant in the morning class. The subject matter had no effect on concentration scores.

The external validity of the study conducted by Raviv and Low (1990) is limited because they did not describe the nature of the PE and science classes, nor did they perform more than one trial to replicate results. Their findings would have been more reliable if they had they performed the experiment over the course of several weeks or months, rather than just providing a snapshot of student concentration. The study's external validity is further threatened by the lack of information on study subjects; the authors gave no information on the population studied other than age, and one does not know to what populations the results are generalizable.

Raviv and Low (1990) controlled for teaching style variation by using the same teachers to teach the morning and afternoon subjects. However, the authors did not measure teacher fatigue over the course of the school day, and this could strongly impact student concentration during lessons. Some description of the classes, as critiqued above, may have alleviated this concern.

The four studies on the effect of physical activity on tasks of concentration and attention found mixed, but generally positive, connections between the two. Pellegrini et al. (1995) found students' attention improved after recess and if recess was not delayed, and replicated these findings in two more experiments. Caterino and Polak (1999) found that 4th graders who exercised immediately before a concentration test significantly outperformed their 4th grade counterparts, but no differences were found for 2nd and 3rd graders. Budde et al. (2008) found a group that took a concentration test immediately following 10 minutes of coordinated exercise showed significantly higher improvement on the test compared with a group that engaged in 10 minutes of regular, non-coordinated exercise. Raviv and Low (1990) found that mean concentration scores were elevated towards the end of classes PE classes. The next section of this review examines the link between physical activity and measures of academic achievement.

Physical Activity and Academic Achievement

Four studies in this review examined the relationship between academic achievement and time spent on physical activity. Tremblay et al. (2000) studied the relationship between physical activity levels and the standardized test results of grade 6 students in New Brunswick, Canada and found that physical activity levels were negatively related to academic achievement. Dollman, Boshoff, and Dodd (2006) asked whether curriculum time committed to PE predicted standardized reading and math scores for South Australian primary school students and found that the amount of time spent in PE did not predict test results. Carlson et al. (2008) looked at the relationship between reading and math achievement of K-5 students and the amount of time the students spent on physical activity and found both positive significant and non-significant associations between the two, depending on grade level and gender subgroups. Lastly, Tremarche, Robinson, and Graham (2007) looked at the standardized English and math scores of two 4th grade classes from two different Massachusetts schools with different amounts of scheduled PE time and found significant differences between the two schools in English test scores, but no significant differences on math scores.

Tremblay et al. (2000) found a negative relationship between physical activity and standardized test results of 5,146 grade 6 students in New Brunswick, Canada. Data was taken from the Elementary School Climate Study (ESCS), a questionnaire administered in spring 1996 to all grade 6 students in New Brunswick asking about the academic and social climates of classrooms and schools, family background, SES, participation in physical activities, and several affective measures including self-esteem. BMI and a general measure of self-esteem were constructed from student responses and merged with standardized test scores in reading and math. Several multiple regression analyses were performed: physical activity on the four covariates of being female, SES, number of siblings, and number of parents; self-esteem on the four covariates, with and without physical activity in the model; and mathematics, reading, and BMI on the four covariates and physical activity, with and without self-esteem in the model. BMI was dichotomized and a series of logistic regression models were fitted to the data.

Tremblay et al. (2000) found that physical activity levels were significantly related to self-esteem ($p < 0.001$). Physical activity levels were negatively related to achievement, but effect sizes were non-significant; a one-unit increase on the activity scale was associated with a reduction in test scores of 2-3% of a standard deviation. The authors concluded that the negative effect associated with a dramatic increase in physical activity amounts to about one to two weeks of additional academic instruction.

A strength of the study conducted by Tremblay et al. (2000) is that they controlled for variables found to affect academic achievement, namely SES and parent education. The authors noted that students who did not have survey data had test scores slightly lower than those who did not. However, the difference was only 6% of a standard deviation and they suspected that non-response bias would not substantially influence their regression estimates. This open and objective analysis gives the study more credibility. However, physical activity was self-reported by students, who were asked about activity that made them breathe hard in the previous week; this seems an unreliable measure of activity. The findings of Tremblay et al. can be generalized to rural populations similar to that of New Brunswick's. The province has fewer than a million people in it and its largest city fewer than 100,000. However, the cross-sectional design used limits discussion of findings to a statement of relationships, rather than cause and effect patterns.

Dollman et al. (2006) surveyed all 450 South Australian government primary schools to examine the connection between time spent in PE and schools' standardized test scores. Only 117 schools returned the survey, but the authors reported the schools are representative of all South Australian primary schools regarding measures of SES, percentage of students with a non-English-speaking background (NESB), school size, and school location. Average PE time for the schools was 133.7 minutes. Dollman et al. did not report whether the 117 schools were representative of state academic achievement, as measured by the State Literacy and Numeracy Test (SLNT). SLNT results were collected at the school level for Years 3, 5, and 7 (roughly equivalent to grades in United States public schools) and weighted by number of students in each level.

Dollman et al. (2006) conducted multiple regression analyses to estimate the connection between time spent in PE and SLNT scores, controlling for SES, NESB, the schools' staff age profiles, and the years of staff PE training; information on the latter two variables was collected via the surveys. The results of the regression analyses showed no significant connections between time spent in PE and SLNT scores for either literacy (*p* = .687) or numeracy (*p* = .876).

Despite the representativeness of the sample obtained by Dollman et al. (2006) with certain individual- and school-level demographic indicators, its internal validity is limited by two factors. First, survey respondents may have been more likely to have stronger PE programs than non-respondents. Second, schools were not surveyed about the resources allocated to PE, nor specific PE activities performed. However, Dollman et al. controlled for staff age and PE expertise, the results appear to be objectively reported, and study limitations are admitted to, strengthening the study's transparency and objectivity.

Carlson et al. (2008) used data from the Early Childhood Longitudinal Study, Kindergarten Class of 1998 to 1999 (ECLS-K), a nationally representative sample, to look at the association between time spent on PE and reading and math achievement. Student data was analyzed at five time points: fall and spring of Kindergarten, and spring of 1st, 3rd, and 5th grades. Time spent on PE was collected from teacher reports. Reading and math achievement was calculated using Item Response Theory (IRT) scale scores reported in the ECLS-K dataset; IRT scales are used in longitudinal data to account for test variations over time. Maximum scores are 186 for reading and 153 for math.

Demographic information, including race/ethnicity, family income, mother's education, and whether the child was enrolled in half- or full-day kindergarten was collected from phone calls to parents. The authors excluded from the final analyses all students who were missing data on any of these variables, and those who were not in the expected grade level at all time points. The final sample consisted of 5,316 students and was 52.1% female, 69.2% non-Hispanic White, 11.1% Hispanic, 8.7% non-Hispanic Black, and 1.9% other (Asian, Pacific Islander, American Indian, Alaska Native, and multi-race). About 70% of the sample had mothers who completed at least some college. About 36% of the sample had family incomes greater than $75,00 and 14.9% had family incomes less than $25,00. About 45% of the sample was enrolled in half-day kindergarten.

Carlson et al. (2008) grouped students into low (0 – 35 minutes), medium (36 – 69 minutes) or high (70 or more minutes) PE groups and analyzed IRT scale scores for math and reading for each grade level by PE category for boys and girls. Significance was determined at the .05 level. Girls in all grades who were in the low PE category had the lowest IRT scores for math and reading, though the difference was only significant in kindergarten and 1st grade and only significant between the low and medium PE categories for kindergarten math. In 5th grade, differences were significant for reading only. No significant associations were found for boys. Carlson et al. found similar results when they used a multivariate linear regression models to test the longitudinal association between PE and IRT scores for math and reading, stratified by gender. The dependent variable was the IRT score at each grade level. Independent variables included PE, grade, baseline IRT, time from baseline, mother's education, family income, race/ethnicity, and half- versus full-day kindergarten.

Only those in the high PE category showed a small but significant benefit for both reading and math scores (reading: 2.4, CI – 1.0-3.9; math: 1.5, CI – .4, 2.7), compared with the low PE category.

Carlson et al. (2008) admit several threats to the internal validity of their study. Most prominently, the final sample for analyses excluded students who were not in their expected grade level and students who did not have data at all of the analysis points on time spent in PE. These criteria excluded over 3,500 students from the final analyses, more than a third of the original sample and a group the authors suspected was generally less White and having lower family incomes. Additionally, the authors do not compare the demographics of their final analysis sample with those of the original sample or other national data; though the original ECLS-K sample was nationally representative, there is no way of knowing if the same can be said of the sample used by Carlson et al. Another weakness of this study is that there was no assessment of the accuracy of teacher-reported data on time spent in PE, nor any way of knowing whether students were active during PE time. Despite these flaws, the study conducted by Carlson et al. was objective and transparent, and their conclusions that time spent on PE has little to no significant effect on the reading and mathematics scores of elementary school students were reasonable based on the data.

Tremarche et al. (2007) looked at the standardized English and math scores of 311 9-11 year old students in two 4th grade classes from two different southeastern Massachusetts schools with different amounts of scheduled PE time. School 1 had approximately 600 students in grades 4-6 (207 4th grade students) and devoted about 28 hours to PE per school year in 2001. School 2 had approximately 1100 students in K-8 (104 4th graders) and devoted about 56 hours to PE in the same year. The two schools had similar PE curriculums in terms of skill foci and activities, and both PE teachers had about 14 years of teaching experience. Students in the schools were required to take Massachusetts standardized English and Language Arts (ELA) and mathematics tests in April and May. Performance scale scores were reported for achievement tests and independent t-tests were used to determine significant differences, set at $p < .05$, between the two groups. A significant difference was found between the two groups' mean ELA test scores in favor of the higher PE school ($t(309) = 1.645$, $p < .05$). No significant difference was found for math scores ($t(309) = .58$, $p > .05$).

A strength of the study conducted by Tremarche et al. (2007) is that the two schools had similar PE curricula and PE teachers with a similar amount of experience. The researchers also conducted a survey that determined the amount of time spent in sports outside of school, as well as the amount of tutoring students received; the survey indicated similar results for the two schools. However, less than half the sample returned surveys, and self-reports of these measures may not be reliable. There are other severe threats to the study's internal validity. Tremarche et al. used a convenience sample that was shown to be very different in terms of ethnicity, SES, dropout and attendance rates, per pupil expenditures, and teacher salaries; School 2 fared better on all economic indicators. Furthermore, as with other studies, Tremarche et al. fail to account for differences in curriculum and other academic variables, other than outside tutoring, that may have affected test scores in the schools.

The four studies in this section found mixed results regarding the relationship between time spent on physical activity and academic achievement. Tremblay et al. (2000), found a non-significant negative relationship between the two. Dollman et al. (2006), found no relationship between time spent in PE and test results. Carlson et al. (2008) found both positive significant and non-significant associations between time spent on physical activity and reading and math achievement. Lastly, Tremarche et al. (2007) found that one school that devoted more time to PE had significantly higher English test scores than a school that devoted less time to PE, but there were no differences regarding the two schools' math scores. The next section reviews the effect of physical exercise curricula changes on achievement test results.

Physical Exercise Curricula Changes and Achievement Test Results

Ten studies in this review examined the effect of physical exercise curricula changes on academic achievement test results. Four of these studies looked at comprehensive, school-wide programs. Sallis et al. (1999) studied the effects of a multiyear comprehensive PE program on the achievement test results of elementary students in California and found that the test scores of intervention students experienced significantly less decline than those of control students. Donnelly et al. (2009) studied an intervention that incorporated physical activity into the regular math and reading lessons of 2nd and 3rd grade students in Kansas and found that those in the intervention condition showed significant improvements in achievement test scores compared to control students. Hollar et al. (2010), in a controlled study of the effects of a comprehensive nutrition and physical activity program on Florida elementary students' achievement test scores, found significant improvement for the intervention group on math scores but no significant differences on reading scores. Ahamed et al. (2007) examined how an increase in a wide range of fitness-related activities affected the achievement test scores of 4th and 5th grade Canadian students and found that academic performance was not affected by the amount of time spent on exercise.

Four more studies looked at the effect of increasing PE. Ericsson (2008) examined the effects of daily PE versus twice a week PE on the academic achievement and attention of Swedish elementary school students and found significant differences in favor of the daily PE group on achievement test scores, but no difference in attention. Corder (1966) studied the effects of one hour daily PE on IQ test results of 8 educable mentally retarded boys and found that IQ scores improved at posttest for the PE group; there was evidence of the Hawthorne effect, as the PE group scores were not significantly better than a group of students who observed the PE classes. McCormick, Schnobrich, Footlik, and Poetker (1968) looked at reading test results for Illinois elementary students in a perceptual-motor training program, additional regular PE time, and no additional PE time and found statistically significant gains from pretest to posttest for the perceptual-motor training group over the other two. Pollatschek and Hagan (1989) studied the effects of daily PE versus twice a week PE on the reading and math achievement of elementary students in Scotland and found no significant differences between the two groups from pretest to posttest.

The two remaining studies looked at the effects of other types of modifications to the regular physical activity of students. Coe et al. (2006) examined the effects of taking PE in either the first or second semester of school and found that standardized test scores of 6th grade students in Michigan were unaffected by the timing of PE enrollment. McNaughten and Gabbard (1993) studied the effects of different amounts of walking at different times of day on the math test scores of 6th graders and found that students who walked for a longer duration performed significantly better than students who walked less or not at all.

Comprehensive School Wide Physical Education Programs

Sallis et al. (1999) studied the effects of a two-year comprehensive PE program on the achievement test results of two cohorts of students from seven public elementary schools in affluent suburban southern California. The seven schools were stratified by percentage of ethnic minorities and randomly assigned to one of three conditions: a Specialist condition (SC) in which trained specialists implemented the intervention; a Trained Teacher condition (TT) in which regular classroom teachers were trained to provide the intervention; and a Control condition (CC), in which students received regular PE lessons but no additional interventions. Participants were 1,538 students who completed a baseline survey. There were no significant differences in gender or ethnic distribution by condition, though there was a significant difference in age by condition ($p < 0.01$).

The Sports, Play, and Active Recreation for Kids (SPARK) program was designed to increase physical activity in school through a three-days-per-week PE program of 30-minute lessons and out of school through weekly 30-minute classroom lessons teaching students behavior change skills. All schools, including controls, were provided with equipment and supplies to implement the SPARK PE program. Family involvement was strongly encouraged.

Academic achievement was measured using Metropolitan Achievement Tests (MAT6 and MAT7), norm-referenced tests that provide scores for reading, math, language, and a composite score. Cohort I was tested in the spring of 2nd and 5th grades. Cohort II was tested in spring of 2nd grade with MAT6 and then fall of 6th grade with MAT7. Scores from the MAT6 and MAT7 are not directly comparable, so cohorts were analyzed separately. In the first stage of analysis, Sallis et al. (1999) regressed posttest achievement scores on baseline scores, sex, dummy variables for group, and all possible interaction terms. Baseline scores among groups were not significantly different for the composite achievement score. All baseline achievement scores were significantly higher than the national average. A significant sex by group interaction was found in one test and was therefore analyzed with analysis of variance, with baseline achievement as the covariate and group as the grouping variable.

There were no significant differences among groups in Cohort I for math or composite scores, but there was significantly less decline for the TT group than for the CC group for language scores ($F(2, 270) = 3.37$, $p = .04$), and a significant increase in reading scores for SC group over the CC group (F(2, 182)

= 4.29, p = .02). For Cohort II, the TT group had significantly less decline than the other two groups on the composite score ($F(2, 423) = 7.23$, p = .001) and reading ($F(2, 423) = 8.36$, p = .001). The SC group saw a significantly greater decline on the language score than other two groups ($F(2, 423) = 5.8$, p = .004). Sallis et al. concluded that SPARK had no detrimental effect on standardized achievement scores, and in many cases mitigated declines in scores.

Sallis et al. (1999) were very thorough and objective in their analyses. The major threat to the study's internal validity was an extremely high attrition rate, about 50%. The authors attributed much of the loss to missing data from one entire school though they do not give specific numbers; this was a dubious claim in light of the fact that one group, the SC group for Cohort I, is noticeable smaller than the other five groups, but only by about 100 students, a small percentage of students lost to attrition. The authors also cited as a reason for high attrition the necessity of complete student records, and that this was difficult to obtain in light of parent surveys and physical activity monitors. However, Sallis et al. are for the most part forthcoming about the issue, and attempted to control for it by comparing missing students with those who had data; attrition did not vary by condition, and students with achievement test data were only significantly different than those without by being an average of .1 years older ($p < .01$) and with higher levels of parent education ($p < .02$).

The study has limited external validity due to its use of a generally affluent, high achieving population, and by the presumably high costs of the intervention in terms of resources and time. A strength of the study is that it maintained good program fidelity, with high levels of training in implementing SPARK. Additionally, the fact that control schools were provided the same resources as intervention schools made it more likely that changes in achievement test scores were the result of the actual intervention rather than money and energy put into resources. However, other aspects of school curricula were not controlled and may have played a large role in achievement over the course of an academic year.

Donnelly et al. (2009) conducted a three-year cluster randomized trial that studied the effects on health and academic outcomes of a comprehensive PE curriculum in Kansas. Physical Activity Across the Curriculum (PAAC), modeled after the TAKE 10! intervention, incorporates physical activity into academic classroom lessons. PAAC schools sought to conduct such lessons for 90 minutes each week. Combined with the 60 minutes per week of regular PE, total PE time would be 150 minutes per week and in line with the recommendations of Healthy People 2010 (Donnelly et al., 2009). Twenty-six schools were originally randomized, stratified by school size and urban versus rural location, to either the PAAC intervention or to a control; one school closed during the study and another refused to serve as a control school, leaving 24 schools included in the analyses.

The study began when students were in 2nd and 3rd grades and ended when the same students were in 4th and 5th grades. The primary outcome reported by Donnelly et al. was changes in BMI from pre- to posttest. Secondary outcomes, including academic achievement, were measured in a subsample of students (n=575; DuBose et al., 2008). Donnelly et al. reported significant ($p < .01$) gains in academic achievement, as measured by the Wechsler Individual Achievement Test-2nd Edition (WIAT-2), for the PAAC schools compared to controls. The WIAT-2, Donnelly et al. reported, has excellent inter-rater reliability, internal consistency, and test-retest stability. Scores for these measures range from .89 to .98, though the authors do not report which methods were used to obtain these scores. Additionally, the test-rest score is cited for children of six to nine years of age; one would expect children in the subsample, in 4th and 5th grades at posttest, to be nine to eleven years old, thus reducing the reliability measures used. Donnelly et al. reported the validity of the WIAT-2 through its correlation with other measures of academic achievement; these scores ranged from .29 to .89 and the measurement methods were again without citations.

The most troubling aspect of the academic achievement gains of PAAC children reported by Donnelly et al. (2009) is that no baseline or follow-up measures are shown. Donnelly et al. simply reported the change in scores from baseline to three years in PAAC and control schools without discussing how the change scores were calculated, nor how significance was determined. Furthermore, scores are reported for 203 students, not the 575 students that were reported to comprise the subsample; no explanation is given for the discrepancy in either Donnelly et al. (2009), nor a secondary paper (DuBose et al., 2008).

The PAAC intervention is noted as an inexpensive and easy way to incorporate physical activity into regular classroom instruction. Donnelly et al. (2009) reported the schools used in the analyses as representative of the diversity found in public education around the country, though again they do not cite supporting numbers in their publication. If one can take their analyses at face value, then the external validity of the study would be quite good. Further studies replicating the results found by Donnelly et al. would lend reliability to their findings; until then, given the threats to the internal validity of the study regarding measures of academic achievement, the results found by Donnelly et al. should be interpreted with caution.

Hollar et al. (2010) studied the effects of two years of Healthier Options for Public Schoolchildren (HOPS), a school-wide nutrition and physical activity program, on the test scores of students in five elementary schools in Osceola, Florida. HOPS focused on making school-provided meals higher in fiber and lower in fat content, providing education and awareness materials to students and parents on the benefits of good nutrition and physical activity, and encouraging schools and teachers to promote physical activity in both daily classroom content area lessons and during recess. Hollar et al. compared results for four intervention schools and one control school on the Florida Comprehensive Achievement Test (FCAT), a state-mandated standardized test. The authors used for their final sample the 1,197 students who qualified for the National School Lunch Program (NSLP), in order to more accurately assess the effects of the altered school meal menus. There was a significant difference between the 974 HOPS students and 199 controls with regard to the percentage of White students in either group, but no significant differences regarding other races/ethnicities, gender, or weight; because the entire sample comprised students in the NSLP, the two groups were measured equivalent on SES.

Hollar et al. (2010) used a two-stage approach to present their findings, first looking at the individual level and then at the school, to account for variation between schools while limited by only having five schools in the analysis. The authors analyzed the individual results to create school-specific means, which in turn were analyzed and adjusted for covariates. Hollar et al. present frequency statistics and test them for associations using Chi-square analyses to look at differences in groups according to demographics. They use repeated measures analyses to look at change in performance over time on the FCAT scores. Hollar et al. found significant differences between the HOPS and control students on FCAT scores over the two-year intervention for math ($p < .001$), but not for reading ($p = .08$).

There are several threats to both the internal and external validity of the study conducted by Hollar et al. (2010). The researchers noted the physical activity component as an important part of the HOPS program but did not measure physical activity performed at either the school or individual level. The authors also noted that physical activity was given more weight in year two of the intervention, but did not perform significance tests to evaluate differences between HOPS and control schools by year. The study was further limited by the use of just one control school. While the use of only those students eligible for free and reduced-price meals ensured the nutritional aspect of HOPS would be more accurately assessed, and while this is a population present in nearly every area of the country, findings based on this one subgroup are not generalizable to other populations and their external validity is limited. Lastly, only 350 students were used to analyze changes and differences in FCAT scores. Hollar et al. note that only those participants who had data at all the time points were included in the analyses, but do not mention why only 350 out of 1,197 students had complete data.

Ahamed et al. (2007) conducted a cluster randomized controlled trial that looked at the effects of Action Schools! BC (AS! BC) on academic achievement. Seven experimental and three control schools were stratified by size (greater or less than 300 students) and location (Vancouver or Richmond, British Columbia) to account for different school organizational structures and ethnic demographics. Schools were then randomized into one of two intervention conditions (INT), which differed in the amount of outside facilitation provided to the school, or to the usual practice control group (UP). INT schools, collapsed into one group for the final analyses, were provided with training, physical activity resources, and facilitation, and were expected to require 15 extra minutes of physical activities each day; combined with the regular 80 minutes of PE per week, these schools were expected to deliver at least 150 minutes of physical activity per week. Teachers monitored exercise through daily logs. UP schools received none of the services INT schools did, but UP teachers were instructed in the completion of the daily exercise logs.

Students' leisure time physical activity was measured with the physical activity Questionnaire for Children (PAQ-C), a self-assessment tool that computes a child's fitness score on a scale from one to five. Although this is a self-report measure of activity, Ahamed et al. (2007) cited a report detailing the PAQ-C's validity, and noted that it has a test-retest reliability coefficient of r=.75 for boys and r=.82 for girls. Academic achievement was measured with the standardized Canadian Achievement Test (CAT-3), a grade-specific test that the authors report, "does not contain any bias with respect to age, gender, or ethnicity" (Ahamed et al., 2007, p. 373). One UP school chose not to submit CAT-3 test results for scoring, and one INT school administered the wrong test at follow-up. Seventy-two students either moved or were absent during the follow-up testing date, leaving a total of six INT schools (n=214) and two UP schools (n=73). Students in both groups averaged 10.2 years of age, were mostly Asian, and had average PAQ-C scores of 2.6 (UP) and 2.7 (INT). Baseline measures of self-reported physical activity were obtained in February and March of 2003. Baseline and follow-up measures of academic achievement were taken in June, 2003 and June, 2004, respectively. It is unclear when and how school wide physical activity was assessed.

Ahamed et al. (2007) used independent t-tests to compare baseline CAT-3 scores, and a mixed linear model to compare follow-up scores between groups and gender. UP students had significantly higher baseline achievement scores than did INT schools (p = .001), but no significant differences at follow-up (-15.3; 95% CI: -41.8, 11.2). Girls and boys performed similarly on the test at follow-up

(p = .27). The authors concluded that the AS! BC program did not negatively impact academic achievement.

The study conducted by Ahamed et al. (2007) has questionable external validity. Though the schools used were reported to be representative of the ethnic diversity found in the greater Vancouver region, the large percentages of Asian and Caucasian students in the study schools are not representative of many other urban areas. Furthermore, AS! BC is a comprehensive school health intervention that relies on extensive materials, training, and outside facilitation; these factors make the intervention a costly one to implement. It is unfortunate that the analyses collapsed the two categories of INT schools, those with extensive outside facilitation and those without, as looking at the two groups separately would be beneficial for schools looking to implement a less expensive intervention.

The cluster randomized controlled design used by Ahamed et al. (2007) ensured a fair amount of internal validity to the study. The authors controlled for differences in baseline academic achievement between the two groups. Additionally, although there was a large difference in the percent of Asian students in the UP (77%) and INT (55%) schools, the authors noted there was no significant difference in baseline achievement scores of Asian and Caucasian students within schools (p = .15), nor was there a significant race-by-group interaction (p = .48). However, Ahamed et al. did not examine the SES of the study schools or other demographic variables besides age and limited ethnic categories (Asian, Caucasian, other).

Attrition was a serious problem in the Ahamed et al. (2007) study: the researchers lost 72 students in addition to two entire schools for their final analyses. Only those subjects who had both baseline and follow-up academic test score data were included in the final analyses, which may have significantly impacted the final results. Additionally, Ahamed and colleagues do not appear completely objective in their analyses. They report that INT schools averaged 183 minutes of physical activity at baseline, while UP schools averaged 140 minutes. They further state that during the intervention, INT teachers delivered an average of 47 minutes more of physical activity than did UP schools. This represents barely a change in activity levels of schools prior to and during the intervention, yet the authors do point this out in their analyses. Furthermore, they state that, “all but three INT schools were within 10% of meeting the guideline of 150 min-wk [of physical activity]” (Ahamed et al., 2007, p. 374). However, this represents half of the INT schools used in the analysis. The discussion of amount of time spent on physical activity is not clear and one wonders if this is done intentionally to obscure the extent to which INT schools were actually changing their activity curriculum.

The four studies that looked at comprehensive school-wide physical activity programs’ effects on academic achievement found mixed results. Sallis et al. (1999) found that achievement test scores of students in a multiyear comprehensive PE program decline less than those of control students. Donnelly et al. (2009) found that students who had physical activity incorporated into regular classroom lessons showed significant improvements in achievement test scores compared to control students. Hollar et al. (2010) found a significant improvement in math, but not reading, scores for students in a comprehensive nutrition and physical activity program. Lastly, Ahamed et al. (2007) found that academic achievement was not affected by participation in a wide range of activities in the school curriculum. The next section of this review looks at increasing the amount of time students spend in PE.

Increased Physical Education and Academic Achievement

Ericsson (2008) examined the effects of different doses of PE on the achievement and attention of Swedish students in years 1 to 3 (similar to grades 1 to 3 in the United States). Year 1 and 2 students attended either traditional PE or more focused sports lessons every day, while year 3 students maintained the usual twice per week PE schedule. Students' motor skills were assessed when they entered the school system and an extra weekly lesson focused specifically on motor skills was provided to students deemed in need of it. Student academic achievement was measured by school assessments of reading development and national tests of students' Swedish and math abilities. Attention was measured with Conners' abbreviated questionnaire. Motor skills were assessed with MUGI, a program of observational assessment during which students complete 16 motor skills tasks. Ericsson cites strong inter-rater reliability of MUGI measurements by school teachers and physicians (.84), internal reliability of assessors (Spearman rank correlation of .75), and test-retest reliability (Spearman rank correlation of .78). Ericsson also cites research on the validity and reliability of the Conners' abbreviated questionnaire. No additional information is given regarding the validity of the internal reading assessments or the Swedish national tests.

Year 1 and 2 students (n=152) were combined in the analysis because there were no significant differences between the two groups on any measures. Ninety-nine students were in the year 3 control group. Ericsson (2008) used non-parametric statistics to analyze the results due to the data's ordinal nature; significance was determined using Cramer's index and values of eta squares. Though students in the intervention group showed improved attention scores in years 1 and 2, there were no significant differences between intervention and control students in year 3; this suggests that the extra PE did not improve student attention scores. Ericsson found significant differences between the two groups on the national tests for both Swedish (Cramer's index of .29) and math (Cramer's index of .21). The differences in achievement test results for students who needed and received extra lessons in motor skills, compared with similar students who did not receive the extra lessons, were also significant, though Ericsson does not give statistics for this claim.

The external validity of the findings presented by Ericsson (2008) is limited by the lack of sample randomization. However, Ericsson cites the sample's baseline measures of motor skills and attention as similar to those of other studies conducted with similar populations. The study was conducted within the practical considerations of a whole-school setting using all students in a grade, lending the findings more external validity than they would otherwise have.

There are also several problems with the study's internal validity. Most importantly, as outcome data was collected in year 3 for each of the years in the sample, there is greater possibility for results to be influenced by other events that took place during the years studied, or the school setting as a whole. Major local events, changes in school policy or routine, or demographic changes in the community could all have affected the study results and would go unnoticed with this study design. On a related note, one cannot assess whether students in each year were similar, as a group, to the others. Ericsson noted that sample characteristics were collected through surveys that were part of a larger study, and that responses indicated the groups had similar attitudes towards physical activity, activity outside of school time, and parents' education and income. One must be cautious, however, at giving these statements too much weight, as Ericsson does not cite the response rate or the representativeness of the responders when talking about the survey. The study conducted by Ericsson does seem to be an objective, valid attempt given the limitations of the data and the nature of a school setting, but further research is needed to confirm the findings' reliability.

Corder (1966) studied 24 "educable mentally retarded boys" (Corder, 1966, p. 358) between the ages of 12 and 16 in Nashville, Tennessee public schools. Two groups of eight boys were taken from one school and randomly assigned to the experimental group or an officials group (observers of the experimental group). Another eight boys were taken from two other schools to serve as the control group. All three groups were closely matched on age and pretest IQ, as measured by the Wechsler Intelligence Scale for Children (WISC), a scale consisting of 12 tests divided into two subgroups: performance and verbal. Scores on the WISC for the sample ranged from 50 to 8. The experimental group went through a series of planned exercises for one hour each school day for four weeks. The officials group attended the daily exercise sessions with the experimental group and, rather than participating in the same activities, kept records of the experimental group's progress on certain tasks. The purpose of the officials group was to measure the Hawthorne Effect – whether or not a change would be observed simply from being watched and encouraged by the researchers. The control group received usual classroom instruction. WISC scores were obtained before and after the intervention.

Corder (1966) used a t-test to determine differences between groups on the WISC and found that those in the experimental group had made significant gains over the control group on both the full WISC scale ($t = 2.65$, $df = 21$) and the verbal scale ($t = 3.67$, $df = 21$). No other significant differences were found among groups on measures of IQ as a function of physical activity. The lack of significant improvement for the experimental group over the officials group indicates the presence of the Hawthorne Effect; however, the lack of significant improvement for the officials group over the control group indicates that the Hawthorne Effect alone could not explain the differences found between the experimental and control groups.

The work of Corder (1966) has little external validity, given the special nature of the population studied, the carefully controlled conditions under which the experimental and officials groups participated in the study, and the amount of time that has elapsed since Corder performed his study. It is difficult to know, based on the study as presented, what sorts of conditions categorize a student as educable and mentally retarded. The study's external validity is further limited by the absence of any girls in the sample, and by the small sample size overall. However, the study was carefully conducted and the work can be generalized to other groups equivalent to that Corder described, groups that are certainly present in public school populations and in need of physical activity interventions. Corder's findings were similar to those of Oliver (1958), and McCormick et al. (1968), lending the study more reliability.

Despite the limited external validity of the study, Corder (1966) was very clear and objective in outlining the intervention and testing procedures, selecting participants, and reporting results. A typical lesson is outlined in detail, giving the reader clear insight into what the experimental and officials groups did. The same psychological examiners and PE teachers administered both the pretests and posttests in all cases but three. The selection and assignment of students to groups was clear, and Corder performed Hartley's test for homogeneity among groups on age and IQ. The resulting F ratios indicated closely matched groups for both age (1.19) and IQ (1.32). However, the fact that age and IQ were the only variables used to establish equivalent groups is a threat to the internal validity of the study. Furthermore, physical activity outside of school hours was not monitored or reported; students in the control or officials groups could have engaged in similar activities as the experimental group. Lastly, it is not clear if the pretests and posttests were given immediately before and after the intervention. This study is applicable to a special population of students and indicates a Hawthorne Effect in students' improvement on IQ measures, but should be interpreted with caution when applying the findings to broader settings.

McCormick et al. (1968) looked at the effects on Lisle, Illinois first grade students' reading test results of three types of physical activity: a perceptual-motor training program (PMTP), additional regular PE time, and no additional PE time. The PMTP consisted of two 45-minute periods a week for seven weeks before school, wherein students learned and performed sequence-based motor skills grounded in motor development research. Students initially did the exercises under commands and later transitioned to self-directing through the sequences. The purpose of the PMTP was to increase students' attention spans. The second group of students met for the same amount of time and at the same time of day, but engaged in regular PE classes with games and activities not focused on motor skill development. The third group of students did not receive any additional PE time.

McCormick et al. (1968) analyzed baseline reading scores, as measured by the Lee-Clark Reading Test, Primer (Form A) using the Kruskal-Wallis one-way analysis of variance by ranks and found differences between groups to be non-significant. The three groups were also of similar age and had similar IQ scores as measured by the California Test of Mental Maturity. McCormick et al. noted that it was suggested they use analyses of variance to test the differences between groups rather than analyze the difference between each group's pretest and posttest. They chose the latter because, they argued, it was more likely to identify real gains that were made, and that may have persisted or broadened over time. McCormick et al. analyzed the differences between each group's score on Form A (pretest) and Form B (posttest) using the Wilcoxen matched-pairs, signed rank test and found that only the PMTP group showed a significant gain in scores ($p < .01$).

There are threats to the internal validity of the study conducted by McCormick (1968). The research question involved the effect of perceptual motor training on the reading achievement of slow or underachieving children, yet the authors do not discuss what they mean by slow or underachieving nor do they give any evidence that their study sample included such students. The authors stated that the randomized groups were equal on gender, age, IQ, and classroom represented (one of the three first grade classrooms in the school); however, no information is given on the SES of the students, the reading instruction they may have received at home, nor their physical abilities at the time of randomization. Another threat to the study's internal validity is the lack of any description of the validity or reliability of the Lee-Clark Reading Test, Primer. It is also unclear if students in the third group received any PE instruction during regular school hours. A strength of the study is that it controls for the Hawthorne Effect by using a group that received extra PE but not the focused intervention.

The biggest threat to the external validity of the study by McCormick et al. (1968) is that the research was conducted almost 50 years ago and probably not generalizable to public school settings today. Another threat to the study's external validity is the extensive training required to implement the PMTP; the authors noted that it takes a full quarter at the graduate level, in addition to practicum fieldwork, to even begin to be acquainted with administering the program. The study's external validity is further limited by its small sample size of 14 students in each group, and by using only first graders. However, the study's external validity is strengthened by its replication of the findings of Corder (1966), which also controlled for the Hawthorne Effect.

Pollatschek and Hagan (1989) studied the effects of daily PE versus twice a week PE on the reading and math achievement of elementary students in Scotland. The daily PE group consisted of 399 primary six and seven (ages 9-12) students from five different schools, while the twice-weekly PE group comprised three other schools chosen based on geographic proximity. The authors report that all schools were of equal SES except one, but do not report any other baseline measures, nor do they report how many students were in the control schools. The PE curriculum usually consisted of 45-60 minute periods of activity, though sometimes activities such as orienteering took place over half a school day.

Reading attainment was measured by the GAPADOL, a Cloze-type test that requires students to fill in words in a given text passage. Math achievement was measured by the Staffordshire Test of Computation (STC), which measures computation accuracy and ranks the student in relation to others of similar age. Pollatschek and Hagan (1989) cite evidence for the measured usefulness of the GAPADOL (Vincent & Cresswell, 1976), but not for the STC. Pre-intervention tests were given at the end of the school year prior to the institution of daily PE, and post-intervention tests given at the end of the following school year. Only students with pre- and post-intervention information were included in analyses. A two-way analysis of variance was conducted using student change scores. Neither a group, sex, nor interaction effect were found on either test at the p = .05 significance level. Similarly, when a two-tailed t-test was conducted on both pre- and post-intervention scores for the two groups, no significant difference was found. The results were the same when the higher SES control school was removed from analyses. The authors do not provide detailed descriptions of the statistics calculated, nor do they provide the ns for either group. They note that change scores for the daily PE group were greater than the regular PE group though again the difference was not significant.

There are several threats to the internal validity of the study by Pollatschek and Hagan (1989). The most pressing of these is the lack of information on the sample used for achievement measures and on the statistics used in the analysis. Achievement was a secondary measure in the study behind fitness measures; however, achievement was measured and results provided, and the lack of detail makes the findings much weaker. Secondly, the authors did not provide any demographic information on any subjects, although they were open in describing their study as using non-equivalent groups. Third, no evidence was cited for the validity or reliability of the GAPADOL, only its “usefulness” (Pollatschek & Hagan, 1976, p. 344), and did not even provide that for the STC; the authors discussed improvements in motor skills for subjects in the intervention and more appropriate tests of motor skill ability could have been used. The study's external validity is bolstered by the practical intervention used – simply increasing school wide PE time. However, given the lack of information on the study's population, it is difficult to determine the results' generalizability.

The four studies that looked at the effect of increasing PE found mixed results. Ericsson (2008) found that a daily PE group's achievement test scores improved, but not their attention. Corder (1966) found that IQ scores improved at posttest for a PE group, although there was evidence of the Hawthorne effect, as the PE group scores were not significantly better than a group of students who observed the PE classes. McCormick et al. (1968) found test scores improved for a group that had perceptual-motor training over a group with increased regular PE time and a group with the same amount of PE time. Pollatschek and Hagan (1989) found no significant differences on reading and math scores for a group with increased PE time versus a group with the regular amount. The next section of this review looks at the effects of having PE at different times of day and different times of the school year.

Effects of Other Physical Activity Changes on Academic Achievement

Coe et al. (2006) conducted their study in a public school in western Michigan in the 2002-03 school year and looked at whether taking PE in the first or second semester of school affected academic achievement. According to regular school policy, students were assigned to a PE class in either the first or second semester of the school year. All 6th grade students were invited to take part in the study, but only 36.8% of students returned the proper forms. Of these

229 students, 15 had missing data. The average age of the 214 students in the final sample was 11.5 years. Fifty-one percent were boys, and the sample was 68% White, 14% Hispanic, 3% Black, 3% Asian, and 12% of other races/ethnicities. The average income of families in the school district was higher than for the state of Michigan, though the authors do not give any data on this.

Coe et al. (2006) measured students' physical activity outside of school using the 3DPAR, which asks students about their activities on the previous three consecutive days. Students were given scores of 1 to 3, with 1 being no activity and 3 being activity that meets Healthy People 2010 guidelines of 150 minutes per week for moderate physical activity and 60 minutes per week of vigorous activity. In-school physical activity was measured using the System for Observing Fitness Instruction Time (SOFIT), which is based on direct observations of PE classes. Academic achievement was based on grades and Terra Nova scores; the Terra Nova is a national standardized test consisting of assessments in reading, math, science, and social studies. Coe et al. measured achievement and physical activity at the end of each semester.

Coe et al. (2006) found that grades and Terra Nova scores were not normally distributed, and therefore compared academic achievement using Kruskall-Wallis analysis. Coe et al. found that students who met the Healthy People 2010 guideline for vigorous activity (60 minutes per week) had significantly better grades in both the first ($X^2 = 1.1$, $df = 2$; $p < .006$) and second ($X^2 = 6.05$, $df = 2$; $p < .049$) semesters. No other significant differences were found for student achievement as a function of either the timing of PE or the amount of observed or reported physical activity.

There are several threats to the internal validity of the study conducted by Coe et al. (2006). First, the authors did not obtain baseline pretest grade data for the sample, making it hard to tell whether students in PE class first semester were affected by activity. Second, the only significant relationship found was based on students' self-reported outside-of-school activity levels; furthermore, the authors do not cite any sources or data that support the validity or reliability of the 3DPAR, only noting that a variation of the test was used in other research conducted by Weston, Petosa, and Pate (1997). Third, about two-thirds of students in the school did not return assent and consent forms, meaning the students who comprised the final sample were self-selected. Finally, Coe et al. note the absence of SES data limits the conclusions that can be drawn from their study, as SES has been shown to correlate with both increased physical activity and higher academic achievement.

The study by Coe et al. (2006) had some strengths as well. First, participating students were randomly assigned to the two PE conditions and the two groups were equivalent on all measures except for BMI. Second, the findings support and are supported by similar conclusions that academic achievement is not affected by increased PE time in school, giving the study some reliability and adding to the research base around the connection between physical activity and academic achievement.

McNaughten and Gabbard (1993) used a Solomon Four-group design to study student performance at different times of day on a math concentration test after walking for 20, 30, or 40 minutes, or not walking at all. The researchers divided 120 6th grade students into four groups of 15 boys and 15 girls each, with a mean age of 11.3 years. Subjects were selected, out of a pool of 151 volunteers, based on grade level math performance according to standardized test scores. The experiment took place over a five week period, with pre and posttests administered to all groups in weeks one and five, and to experimental groups in weeks two thru four. For the test, students were given 90 seconds to complete as much as they could of 40 3rd grade level math problems. Students took the tests at three different time periods: 8:30, 11:50, and 2:20.

McNaughten and Gabbard (1993) used repeated-measures analyses of variance to compare posttest scores of the experimental and control groups and found no significant differences ($p < .05$) on math test performance without preceding exertion. The authors then used repeated-measures factorial analyses of variance to look at the relationships among gender, duration of activity, and time of activity. Students did significantly better on the test after exercise durations of 30 and 40 minutes at 11:50 and 2:20, compared with students who walked 20 minutes or not at all. The authors conclude that 30-40 bouts of moderate exercise help academic performance.

The Solomon Four-group Design used by McNaughten and Gabbard (1993) has several strengths. It allows for comparisons of control and experimental groups, and of pretest-posttest subjects versus those who just took a posttest. Furthermore, the researchers took steps to ensure the internal validity of the study, such as having the experimental subjects practice walking at a moderately intense pace and randomization of subjects to experimental and control groups. However, the study also has some limitations. First, the study sample was not randomly selected, as students volunteered to participate. Second, the authors did not describe student preferences for math or physical activity at baseline, nor any student characteristics except for mean age and gender. Third, the control students and experimental subjects were tested during different weeks, and no mention was made of any differences in scheduling or other circumstances that may have affected the internal or external validity of the results.

The two studies in this section of the review looked at the effects of very different changes in physical activity. Coe et al. (2006) found that students' standardized test scores were unaffected by having PE in the first or second semester of school. McNaughten and Gabbard (1993) found that students who walked for a longer duration during the day performed significantly better on math tests than students who walked less or not at all.

The ten studies in this section of the review all looked at the effects of physical exercise curricula changes on student achievement on tests. The final section of the review examines the effect of physical activity on other areas of school performance.

Physical Activity and Other Areas of School Performance

The remaining six studies in this review looked at other relationships between physical activity and school performance. Hill, Williams, Aucott, Thomson, and Mon-Williams (2011) studied the effect of daily exercise on the cognitive performance of 552 Scottish elementary school children and found that students who exercised performed better than those who didn't in the second week of the study, but not in the first week. Tuckman and Hinkle (1986) studied the effect of running versus regular PE classes on the creativity and planning ability of 4-6th grade students and found that the running group performed significantly better on measures of creativity but not on measures of planning ability. Della Valle et al. (1986) examined whether the learning preferences of urban junior high school students affected their performance on a recall test and found that those who preferred active learning environments performed better on the test after an active lesson than those who preferred passive learning environments, and vice versa. DeBusk and Hellison (1989) studied the effect of a PE program conducted during noon recess on the behaviors of 10 4th grade boys and found no changes in office referrals but positive changes were reported from school staff. Zervas, Danis, and Klissouras (1991) found that students who took a cognitive test, ran for an hour, and then took a posttest performed significantly better on the test, while a control group of students who took the tests without physical activity in the middle did not experience significant improvement. Lastly, Tergerson and King (2002) studied the perceptions of 535

Ohio high school students regarding benefits from exercise and found that many students noted increased self-esteem and energy levels and decreased stress.

Hill et al. (2011) studied the effect of daily exercise on the cognitive performance of 552 8-12 year old Scottish children from nine schools in Aberdeen. Two of the schools were categorized on the Scottish Index of Multiple Deprivation, a five point scale of SES, as least deprived, two as most deprived, and the other five schools in between; 43% of the sample had subclinical Attention Deficit Hyperactivity Disorder (ADHD) symptoms, as measured by the Vanderbilt ADHD Diagnostic Parent-Rating Scale, and 10% were classified as clinical. One school from each level of deprivation was randomly assigned to group A or B. Schools in group A completed classroom-based exercises 30 minutes after lunch for 10-15 minutes each day for one week while schools in group B participated in enjoyable non-physical exercises such as art and games. In the second week, the schools did the opposite. Psychometric tests were conducted in the last 15 minutes of each school day. Tests were given orally and measured students' ability to simultaneously track multiple stimuli. Different versions of the tests were given week by week to reduce practice effects and students violating testing condition rules had results excluded from analysis. Hill et al. set each student's average performance score as a repeated measure and looked at fixed effects of intervention, group, grade, sex, BMI classification, and ADHD classification.

Hill et al. (2011) found significant main effects ($p < .001$) of sex, grade level, ADHD classification, and group on cognitive performance, as well as a significant interaction effect of group and exercise intervention. There were no significant differences between groups in week one, while the exercise group significantly outperformed the non-exercise group in week two (p = .03); the benefits were not moderated by BMI or ADHD classification. The authors suggested that students' familiarity with the test type and conditions, combined with the exercise intervention, produced the improved performance for week two subjects.

The rigorous research design used by Hill et al. (2011), and its implementation, gave the study high internal validity, although no details are given on the validity and reliability of the measures used. The study has high external validity in that the findings replicated an earlier study led by the first author (Hill et al., 2010) that used a similar design, although with a less diverse population and slightly altered cognitive testing. Furthermore, because Hill et al. (2011) used a socioeconomically diverse population with a high percentage of ADHD students, the findings are more generalizable to other broad public school populations. The study's internal validity is weakened by study attrition and exclusion of results that led to differential SES populations in the two groups, with the final study sample representing only about 73% of the original participants. The lack of accounting for possible differences between groups on other activities during the school day, especially academically related activities, is also a weakness of the study.

Tuckman and Hinkle (1986) studied the effects of running on the various academically related psychological behaviors of 154 4-6th grade students.

Participants were randomly assigned to a control group of PE three (6th graders) or five (4-5th graders) times per week or the treatment, which consisted of three 30-minute running sessions per week for 12 weeks; running sessions took place in lieu of PE classes, such that 6th graders had no extra PE and 4-5th graders had two PE periods per week in addition to the running sessions. The researchers speculated that students are not active for much of regular PE periods and that a coordinated running program would be more stimulating. Behavior was measured by teacher responses to the Devereaux Elementary School Behavior Rating Scale (DESBRS), which has a median test-retest reliability coefficient of .81. Creativity was measured by the Alternate Uses (AU) test, reliability estimates of which range from .62 to .85. Self-concept was measured by Piers-Harris Children's Self-Concept Scale (SCSC), which has test-retest reliability of .71 to .77. The Maze Tracing Speed Test (MTST) was used to measure planning ability and visual-motor coordination.

Tuckman and Hinkle (1986) performed analyses of variance on posttest scores for each of 10 measures using pretest scores as covariates to compare groups across sex and grade level. The treatment group performed significantly better than the control group on the measure of creativity ($F(1, 153) = 17.00$, $p < .001$). No significant effects were found for measures of classroom behavior, self-concept, or planning and visual-motor coordination.

The external validity of the study conducted by Tuckman and Hinkle (1986) is strengthened by the researchers' use of entire classes of students and building of the intervention within the regular school curriculum. External validity is further strengthened by the ease with which the intervention could be implemented, and by admission policies of the school that require the student population to reflect the population of the state in terms of race/ethnicity, sex, SES, and intelligence, though it doesn't say how the latter two are measured. The study's external validity is limited by the exclusion from the study of those with physical or mental handicaps, as this is an important population in almost any public school. The fact that Tuckman and Hinkle conducted a randomized study using three entire grades of students is a strength, as is the reporting and reliability of measures used in the analyses. However, the benefits of a random assignment study are counteracted somewhat because the researchers gave no information on the randomization process or the comparability of groups after randomization.

Della Valle et al. (1986) studied whether preferences of 7th grade students for active versus passive learning environments affected their performance on a recall test following lessons in such environments. The Learning Style Inventory (LSI) was administered to all 7th grade students in an urban junior high school of 1,500 students; 50% of the school's students were Black, 30% White, and 20% were Hispanic. The LSI shows grades seven to nine showing reliability of .93 for females and .89 for males, though the methods used to calculate these reliability scores are not reported. Della Valle et al. selected students on the extreme ends of the LSI scale to ensure the students had clear preferences for active or passive learning environments; furthermore, researchers limited the final sample to students with consistency scores of 10. The final experimental sample was 40 students: 17 boys and 23 girls; 20 students preferred passive learning environments and 20 preferred active environments.

In the passive environment, students sat in their seats and word pairs were flashed on a screen at four-second intervals. In the active group, the same students examined 15 different word pairs printed on index cards placed face down on tables around the perimeter of the classroom. Students lifted a card, examined it for four seconds, placed it back face down, and moved to the next card. Immediately following either task, students took a test that consisted of 60 word pairs from which they were instructed to identify those they had seen previously on the screen or around the room. Scores were determined by the number of pairs correctly identified. Della Valle et al. (1986) analyzed correctly identified word pairs with a split-plot analysis of variance that revealed no significant differences between factors other than mobility preference. This significance only occurred when student learning preferences were matched with the learning environment ($F = 4.31$, $df = 1$, $p < .001$), and occurred with each interaction of student preference and environment.

No studies similar to that conducted by Della Valle et al. (1986) have been conducted, and the researchers used word pair recall tests that had not previously been established as reliable and valid. One also wonders if the order in which students took the two tests, after passive versus active environmental conditions, biased the results in some way; there is no discussion of this by Della Valle et al. and thus no way of knowing. For these reasons, the reliability and internal validity of the findings can reasonably be questioned. Nevertheless, the significant interactions found by Della Valle et al. are evidence that students' ability to recall information can be improved by teaching to their preference for learning. Further research is needed to determine if these findings are applicable to other settings and populations.

DeBusk and Hellison (1989) conducted a qualitative study examining the effects of a PE program on problem behaviors of 10 4th grade boys. The program took place during noon recess for three one-hour periods per week for six weeks and included a variety of fitness and sports activities, daily teacher talk on strategies for dealing with conflict, and regular time for student journaling and reflection. Data was collected from several sources: pre- and post-interviews asking open-ended questions about self-responsibility, office referrals for each student, teacher-researcher accounts of the program and feelings throughout the study period, post-program narratives from classroom teachers and recess supervisors, post-program narrative evaluations written by volunteer teaching assistants who were involved in the program throughout, and student daily reflections.

At the end of the six weeks, there were no large changes in the number of office referrals for the participants, but several positive changes were reported from school staff and participants rated the program highly in their journal accounts. Based on the post-interviews and journals, many students seemed to enjoy the program, were more inclined to say they were responsible for themselves, and enjoyed sharing their feelings. The teacher-researcher described slight improvements in all the boys' behavior during regular PE classes; the teacher, teacher assistants, and playground supervisors all noted that the boys' behavior was generally better in the special PE program than in regular PE, and better in regular PE than it was in the classroom or on the playground. Teachers and teacher assistants noted improvements in the boys' self-control and caring, two behaviors the program sought to affect. Teachers also noted positive changes in the boys' self-esteem.

The teachers and investigators in the Debusk and Hellison (1989) study were able to follow the case study model laid out at the beginning of the intervention. The procedure, the goals of the intervention, and the results are clearly laid out; details and descriptions lend the study credibility. Much time has elapsed since the study was conducted, and therefore it is difficult to conclusively establish the study's confirmability. However, it is impossible to include the minutia of 200 pages worth of notes, and the authors clearly state what they did and what they found. The authors noted that their results were similar to those found in another case study conducted by the second author (Hellison, 1978), lending the Debusk and Hellison study dependability. However, the transferability of the model is questioned, as the program procedures followed were extensive and involved the participation of several staff members; further research is needed to determine whether a similar program can be adapted to a bigger population, different timeframes, and a different mix of school personnel. However, the results of the Debusk and Hellison study suggest positive changes can be produced by participation in, and reflection on, a behaviorally focused physical activity program.

Zervas et al. (1991) studied the cognitive performance of students who ran for an hour between taking a cognitive test and a control group tested with no physical activity in between. Nine sets of male monozygotic twins with an average age of 13.1 years were divided into trained and untrained experimental groups. The trained group followed a 25-week physical activity program, in addition to attending regular PE classes two to three times per week, before testing. The untrained and control groups just had the PE classes. Both trained and untrained groups ran for one hour in between the pre- and posttests. The Cognitrone Test, which requires subjects to identify target stimuli, was used to measure outcomes. Groups performed three sets of 50 trials at pretest and three more at posttest. Zervas et al. analyzed the results with a group by test analysis of variance, with dependent variables of right and wrong responses and decision time. They found a significant interaction for correct responses ($F(2, 23) = 4.54$, $p < .02$), as the two experimental groups saw significant gains from pre- to posttest and the control group did not. The authors found no significant differences for the number of incorrect responses. All groups had improved decision times from pre- to posttest and there were no significant differences between any groups; the authors speculated this could be due to participants becoming familiar with the test.

Zervas et al. (1991) use of twin pairs for two different experimental groups strengthened the study's internal validity, especially for comparisons of results between trained and untrained experimental groups. However, no information is given on the control group other than their average age, and one has no idea whether these groups were similar on several important measures, especially baseline cognitive performance. This is less of an issue in light of the researchers' analysis of change scores over an hour of physical activity versus no activity; more information on subjects would have been beneficial nevertheless. Another threat to the study's internal validity is the lack of information given on the validity and reliability of the Cognitrone Test. A strength of the study's external validity is the ease with which it can be implemented in most school settings. However, one does not know how generalizable the findings are given the lack of information on subjects and school and community setting, the fact that all subjects were middle school boys, and the study's small sample size.

Tergerson and King (2002) conducted a qualitative study that looked at what adolescents' viewed as the perceived benefits to physical activity. The sample was drawn from single gender high schools in Cincinnati, Ohio, and consisted of 245 boys and 290 girls. About half the sample were in 9^{th} grade, about a quarter were in 12^{th} grade, and the rest were in 10^{th} and 11^{th} grades. Teachers in randomly selected homerooms gave the surveys at the beginning of a designated class period and read the same instructions for completing the survey. The survey asked students about the benefits they received from physical activity, and the authors grouped similar student responses. Several of these groupings related to academic performance. More than half of student responses indicated that they exercised because it increased their self-esteem and made them feel better about themselves. Similarly, more than half of respondents indicated that exercise increased their energy levels and reduced stress.

The study conducted by Tergerson and King (2002) has questionable credibility given that the authors do not ground their methods in any research on qualitative research methodology; however, it is reasonable to think the study is somewhat credible given the simple nature with which the survey was administered and the methods used to categorize responses. These traits lend the study a high degree of confirmability; the research is open to review and easy to follow. Unfortunately, this review did not find similar studies performed, so the dependability of the study conducted by Tergerson and King is unknown. The biggest factor affecting the integrity of the study is that the sample was drawn from single gender private schools; no research has been done to examine the effect a single gender environment has on perceived benefits of exercise, and this greatly limits the transferability of the study. The predominantly middle class, Caucasian characteristics of the study sample further limit the study's transferability.

The six studies in this section of the review looked at relationships between physical activity and areas of school performance outside the realm of achievement tests, with mixed results. Hill et al. (2011) found that students who exercised performed better on cognitive tests than those who didn't in the second week of the study, but not in the first week. Tuckman and Hinkle (1986) found that students who ran, as opposed to participating in regular PE, performed significantly better on measures of creativity but not on measures of planning ability. Della Valle et al. (1986) found that students who preferred active learning environments performed better on a recall test after an active lesson than those who preferred passive learning environments, and vice versa. DeBusk and Hellison (1989) found that a PE program conducted during noon recess did not have an effect on office referrals but resulted in positive behavior and self-esteem changes as reported by school staff. Zervas et al. (1991) found that students who ran for an hour before taking a cognitive test performed significantly better on the test, while students who took the tests without physical activity in the middle did not experience significant improvement. Lastly, Tergerson and King (2002) found that many students noted increased self-esteem and energy levels and decreased stress as positive benefits resulting from exercise.

Summary

Chapter two presented a critical review of the literature on the connection between physical activity and fitness, and measures of academic performance. Five of the six studies that looked at the relationship between fitness levels and academic performance found positive relationships between the two, while the remaining study found no relationship between fitness and concentration.

The four studies that looked at the effect of physical activity on tasks of concentration and attention found generally positive connections between the two.

The four studies that examined the relationship between academic achievement and time spent on physical activity found mixed results. One study found that physical activity levels were negatively related to academic achievement. Another found that curriculum time committed to PE did not predict test results. A third found both positive significant and non-significant associations between time spent on physical activity and reading and math scores. The fourth study in this section found a school that spent more time on PE performed significantly better than a school with less PE time on English tests, but not significantly better on math tests.

The four studies that looked at the effect of increasing PE found mixed results: one study found improved achievement test scores but not attention; another found improved IQ scores with possibility of the Hawthorne effect; a third found improved test scores; and the fourth found no significant differences for reading and math scores. The four studies that looked at comprehensive school-wide physical activity programs' effects on academic achievement also found mixed results. One study found less decline for students in a comprehensive PE program; another found significant improvement on test scores; a third found significant improvement in math, but not reading; and the fourth found no effect. The two studies that looked at the effects of changes to the regular activity schedule also found mixed results: one study found that test scores were unaffected by the timing of PE class, and the second found that students found that students who walked more during the day performed better on math tests.

The six studies that looked at relationships between physical activity and areas of school performance also had mixed results. One study found that students who exercised performed better on cognitive tests in the first week of a study but not the second. Another found that students who ran performed better on measures of creativity but not planning ability. A third found that students performed better when matched with their preferred learning environment. The fourth study found a PE program did not have an effect on office referrals but resulted in other positive changes. The fifth study found that students performed better on a cognitive test after running. Lastly, the sixth study found that exercise resulted in increases in self-esteem and energy levels and decreased stress.

Chapter three reviews the historical background and rationale for this study laid out in the first chapter and summarizes the research presented in chapter two. Chapter three then delves into the implications of this research for classroom practice in K-12 public education, and presents recommendations for further research.

CHAPTER 3: CONCLUSION

Introduction

Chapter one of this review discussed the history of PE in the United States, in particular the link between emphasis on fitness and American involvement in wars as the military noted recruits' low fitness levels and policymakers viewed the issue as a threat to national security. This emphasis declined in the later part of the 20th century, and remains low today, as attention has shifted to student achievement as measured by standardized tests. Administrators and teachers today must make schedule and curricula decisions balancing the primacy of academics with other student needs. Chapter one also discussed the importance of studying the link between physical activity and academic performance given this increased focus on achievement. If academic performance is negatively affected by curriculum time spent on physical activity, then there should be little room for such activity given policymakers' current goals. If academic performance is unaffected or positively affected by activity, then the policy shift towards less PE and more content-based lessons should be closely examined. This review hopes to inform this discussion by looking at the link between academic performance and physical activity.

Chapter two summarized and analyzed the literature regarding that link, divided into five sections: fitness levels and academic performance, physical activity and academic performance, physical activity and academic achievement, physical exercise curricula changes and achievement test results, and the effects of physical activity on other areas affecting school performance. Findings in each of these sections were mixed. However, very few negative relationships were found between fitness and activity and measures of academic performance; in most cases, either a positive effect or relationship was found, or there were no effects or relationships. Chapter three of this review will summarize the strengths, weaknesses, and trends of the literature reviewed in chapter two; discuss the implications these findings have on classroom practice; and present suggestions for future research in this area.

Summary of Findings

This section of chapter three summarizes the strengths, weaknesses, and trends of the literature reviewed in each section of chapter two. The first section in chapter two looked at fitness levels and academic performance and found mostly positive relationships between the two. Castelli etl al. (2007) conducted the most rigorous study within this section, followed by Chomitz et al. (2009), Grissom (2005), and Hillman et al. (2005). The studies conducted by Dwyer et al. (2001) and Knight & Rizzuto (1993) were the weakest in this section and significantly flawed.

A strength of the literature in this section is the near uniformity in measurement of fitness using the Fitnessgram, a series of fitness tests that has been deemed valid and reliable by the creators (Welk, Morrow, & Falls, 2002); four studies used the Fitnessgram (Castelli et al., 2007; Chomitz et al., 2009; Grissom, 2005; Hillman et al., 2005). The results of this section are further strengthened by the commonality in findings, with five of the six studies showing positive relationships between fitness and academic performance (Castelli et al., 2007; Chomitz et al., 2009; Dwyer et al., 2001; Grissom, 2005; Knight & Rizzuto, 1993) and only one showing no relationship (Hillman et al., 2005). Another strength of the collection of studies in this section is the diverse populations represented, ranging from the entire state of California (Grissom, 2005) to students in an entire school (Chomitz et al., 2009) to small samples of high- and low-fit students (Hillman et al, 2005). Studies also covered all grades in K-12 education. Collectively, the study findings can be applied to a wide range of populations, which is both a strength in broad coverage, but a weakness in lack of consistency from study population to study population.

There were other significant weaknesses to the studies in this section of the review, many of them concerning issues with study populations. Some of the studies experienced issues with attrition (Castelli et al., 2007; Chomitz et al., 2009). Hillman et al. (2005) used a very small sample. Grissom (2005) used all the data available from California, but a large percentage of students had unusable data. The final sample size used in the analyses conducted by Dwyer et al. (2001) was unclear, one of several serious issues threatening the validity and reliability of that study's findings. Similarly, Knight and Rizzuto (1993) gave very little information regarding their sample or methods, decreasing that study's significance as well. Finally, the results presented in this section are framed in terms of relationships and not causality. One cannot conclude that improved fitness causes better academic performance, and one author (Grissom, 2005) notes the conversation could just as easily be framed around the connection between SES and both fitness and academic performance.

The second section of chapter two looked at the relationship between physical activity and academic performance and also found mostly positive relationships. Pellegrini et al. (1995) conducted the strongest of the studies in this section. The remaining three studies were far weaker, with that of Caterino and Polak (1999) the strongest of the three, followed by Budde et al. (2008). Raviv and Low (1990) conducted the weakest study in this group. Looking at the studies in this section as a group, three of the four (Pellegrini et al., 1995; Budde et al., 2008; Raviv & Low, 1990) found improved student concentration and attention after exercise; the fourth (Caterino & Polak, 1999) found the same result in a 4th grade sample but found no differences for 2nd and 3rd graders. A collective strength of these studies is the tight study controls used for coding and inter-reliability (Pellegrini et al., 1995), intervention and testing procedures (Budde et al., 2008), and in classroom designs (Raviv & Low, 1990). The use of similar grade levels for study populations is another strength, though Budde et al. (2008) did not fit this trend. The biggest strength of studies in this section lies in the replications conducted by Pellegrini et al. (1995); taken together, these three experiments have more weight.

Many studies lacked important information on sample characteristics. Raviv and Low (1990) only listed their subjects' ages. Budde et al. (2008) and Caterino and Polak (1999) failed to demonstrate the equivalence of the groups they used; these non-equivalent control designs carry significantly less weight than do randomization processes that result in equivalent groups.

The third section of chapter two looked at the relationship between physical activity and academic achievement. Tremblay et al. (2000) conducted the most rigorous study among those in this section, followed by Dollman et al. (2006), Carlson et al. (2008), and Tremarche et al. (2007). All four studies had significant weaknesses, especially regarding data collection. Physical activity in the Tremblay et al. (2000) study was self-reported through a survey that asked how often participants breathed hard in the previous week, not a very reliable measure; furthermore, the study population was very rural, and not well generalizable to broader, more diverse populations. Dollman et al. (2006) survey responders were representative of all primary schools in South Australia regarding SES and time spent in PE, but one cannot know the nature of the PE classes, nor were responders known to be representative academically. Carlson et al. (2008) lost more than a third of their subjects to attrition, and did not assess their comparability to those retained in the study. Likewise, Tremarche et al. (2007) experienced large non-response on their surveys, and studied populations with very different SES and per pupil expenditure data. Furthermore, these studies, again, looked at correlational, rather than causal, relationships between physical activity and test scores. In light of study flaws, even these relationships must be viewed with skepticism. Nonetheless, taken together, the studies add to a large body of literature on the positive connection between physical activity and achievement among elementary school students; the one negative relationship found was non-significant.

The ten studies that looked at changes to school exercise curricula were divided into three categories: comprehensive school wide PE programs, increased PE and academic achievement, and effects of other physical activity changes on academic achievement.

The four studies that looked at the effects of school wide PE programs found mixed results. Sallis et al. (1999) conducted the strongest of the studies in this section, followed by Donnelly et al. (2009) and Hollar et al. (2010); the study conducted by Ahamed et al. (2007) was the weakest. All four studies shared some severe threats to both internal and external validity. Three of the four studies experienced attrition rates greater than 50%, and the fourth lost a large percentage of subjects as well. Important information regarding baseline and follow-up measures (Donnelly et al., 2009), and physical activity (Hollar et al., 2010) was also missing from studies. Additionally, while Sallis et al. (1999) appeared objective in their discussion of the intervention studied and loss of subjects, Ahamed et al. (2007) did not in their reporting of differences between groups. These threats to study validity make this body of research weaker than it otherwise would have been, given the strong experimental study designs used.

Furthermore, comprehensive school wide programs are generally very expensive to implement. All four of these interventions required extensive training and materials to implement. However, the external validity of the collective studies is strengthened by the use of high SES students (Sallis et al., 1999), low SES students (Hollar et al., 2010), and mixed SES populations

(Ahamed et al., 2007; Donnelly et al., 2009), all of which involved elementary school students.

The four studies that looked at the effects of increasing PE on academic achievement found mixed results, with authors reporting either positive effects or no effects. Ericsson (2008) conducted the strongest study among those in this section, followed by Corder (1966), McCormick et al. (1968), and Pollatschek and Hagan (1989). The external validity of all four studies in this section is highly questionable. Two of the studies (Corder, 1966; McCormick et al., 1968) were conducted well over 40 years ago with very small sample sizes; Corder's (1966) sample was limited to mentally challenged boys. The external validity of these two studies is strengthened somewhat by their similar findings, including evidence of the Hawthorne Effect, in which changes in outcomes are at least partly attributed to simply being involved in a study. The other two studies in this section (Pollatschek & Hagan, 1989; Ericsson, 2008) were conducted in northern European countries, making their generalizability to United States public education settings more questionable.

Even more troubling are problems with the implementation of the study designs: Pollatschek and Hagan (1989) do not report information for the sample used to measure academic achievement, and Ericsson (2008) does not report on the comparability of the groups used in the study. Nevertheless, these four studies of increased PE time in elementary school add to the body of literature supporting positive or null effects of increased PE time on academic achievement.

Two studies in this subsection of the review looked at the effects of very different changes in physical activity. Coe et al. (2006) conducted the more rigorous of the two studies, and found that test scores were unaffected by having PE in the first or second semester of school. McNaughten and Gabbard (1993) found students who walked for a longer duration during the day performed significantly better on math tests. The rigorous experimental study designs used in both studies give them more weight, though each also comes with problems. Coe et al. randomly assigned students to each condition, but a small percentage of students returned the initial survey, no baseline data was given on the subjects, and the findings were based on self-reported outcomes. The Solomon Four Group Design used by McNaughten and Gabbard (1993) allowed for comparisons of control and experimental groups, and of pretest-posttest subjects versus those who just took a posttest, but students self-selected into this study as well, and students' subject preferences could have biased the results.

The ten studies in this section of the review all have serious design or implementation flaws; however, taken together, the studies have more weight given their consistent findings of positive or null effects of physical activity on academic achievement.

The six studies in the fifth section of chapter two looked at the effects of physical activity on other areas of school performance. Hill et al. (2011) conducted the most rigorous of these studies, followed by Tuckman and Hinkle (1986), Della Valle et al. (1986), DeBusk and Hellison (1986), Zervas et al., (1991), and Tergerson and King (2002).

These studies generally found positive relationships between physical activity and measures relating to academic performance, but are hard to categorize given the variety of interventions and methods used. However, statements can be made regarding common strengths and weaknesses. Populations studied were similar, with five of the six looking at students in 4th thru 7th grades. Three studies (Hill et al., 2011; Tuckman & Hinkle, 1986; Zervas et al., 1991) randomly assigned subjects to intervention and control conditions, but all three studies were missing important information on the study sample or equivalence of groups. Hill et al. (2011) and Tuckman and Hinkle (1986) used samples that were generalizable to other common public school populations, while others (DeBusk & Hellison, 1989; Della Valle et al., 1986; Tergerson & King, 2002; Zervas et al., 1991) used either small or non-diverse sample sizes.

The strengths and weaknesses of these studies were discussed in chapter two; though they collectively support the notion that various forms of physical activity benefit various populations, the studies in this section are more remarkable for their uniqueness than they are for their similarity.

Implications for Teaching

This review summarized and analyzed the literature on connections between physical fitness and activity and academic performance in order to inform classroom practice. The collected body of research has several relevant implications to teaching in K-12 schools.

The research suggests that classroom lessons should not extend for too long without breaks for both brain and body. Pellegrini et al. (1995) revealed the necessity for breaks through their finding, replicated in two more experiments, that student attention was higher when classroom time between the beginning of the day and recess was shortened. A further benefit of these breaks is that students' attention was also higher at the end of recess compared to the beginning. Similar findings by Budde et al. (2008), Caterino and Polak (1999), and Raviv and Low (1990), suggest that student concentration can be improved by exercising prior to classroom-based lessons. Zervas et al. (1991) suggested these benefits extend to general cognitive ability, as opposed to specific measures of attention and concentration. Classroom teachers could better hold their students' attention if they make sure continuous sedentary time does not last too long. Implicit in this is that teachers should attempt to capitalize on this increased post-exercise attention and concentration by focusing on key points of instruction immediately after students exercise.

Teaching exercise through intentional, sequenced development of motor skills also appears to enhance academic performance. This is different from simply releasing students from their chairs; it requires careful planning and execution. Both McCormick et al. (1968) and, much later, Ericsson (2005) found that specifically increasing student development of motor skills through sequenced activities resulted in higher academic performance of those students.

Della Valle et al. (1986) found that students performed better on a recall test when they were matched to their preferred learning environment, either passive or active. Teachers can take advantage of this finding by varying their instruction such that both passive and active learning environments are used in lessons. Teachers may also be more successful if they give students options in completing lessons and assessment tasks.

Inconsistent results of studies on comprehensive school wide programs suggest that schools and classrooms should hesitate to institute such programs. The programs studied were also, for the most part, quite expensive to implement, with significant training hours, consultation, and materials. These programs might be worth the cost if they were shown to improve student achievement, but this is not the case. The one exception, in both expense and results, was the program studied by Donnelly et al. (2009). Students in this program took short exercise breaks at their desks during regular classroom lessons. Donnelly et al. (2009) reported the program to be both inexpensive and effective in increasing student achievement. It is worth repeating, however, that the study had serious threats to its internal validity and results should be interpreted with caution.

Other studies also found benefits for short bouts of exercise (Hill et al., 2011; Budde et al., 2008). These findings suggest that teachers can improve student performance with short exercise breaks. These could be conducted in between lessons, or in the middle of longer lessons.

Findings that physical activity enhances self-esteem (DeBusk & Hellison, 1989; Tergerson & King, 2002; Tremblay et al., 2000) are also important for teachers. The pursuit of self-esteem as a goal can be detrimental to learning (Crocker & Park, 2004); however, it can contribute to academic performance when it is a byproduct of some other goal or activity (Shepherd, 1996; Tremblay et al., 2000). Benefits of physical activity to academic performance are further supported by findings that physical activity can reduce disruptive behaviors (DeBusk & Hellison, 1989) and stress (Tergerson & King, 2002). Teachers can apply this in classrooms during stressful times such as long testing days, and in situations where students' may excel in running, sports, or other active situations. Again, it is important to remember that increases in self-esteem should not be the goal, but a byproduct of the activity.

Findings regarding increased PE time are also relevant for elementary school classroom teachers. These teachers typically have greater control over daily schedules. As schools cut back or eliminate PE from regularly scheduled curricula, elementary school teachers must choose whether to include it or not. Ericsson (2008), Corder (1966), and McCormick et al. (1968) found that increasing PE time improved academic performance.

Numerous studies in this review concluded that there are no correlational or causal links between physical activity and academic performance. In contrast to the studies mentioned above, Pollatschek and Hagan (1989) found no benefits of increased PE time. Dollman et al. (2006) found that PE time did not predict achievement. Carlson et al. (2008) and Tremarche et al. (2007) found mixed, but mostly non-significant connections between activity and academic performance. However, the only study that found negative associations (Tremblay et al., 2000) found the correlation to be non-significant. The designs and execution of this research were lacking in many regards. However, taken as a whole, it suggests that maintaining or increasing physical activity does not harm academic performance. Extensive research has shown that exercise increases both physical and mental health (U.S. Department of Health and Human Services, 2008). The enormous costs to the public associated with health care in the United States (Seldon & Sing, 2008) directly impact schools, as these two institutions compete for public dollars. While the causal link between exercise and academic performance should be explored in greater detail, research suggests the overall benefits of providing exercise in school-based settings far outweigh the costs of not doing so.

Suggestions for Further Research

The body of research on the correlational and causal connections between physical activity and academic performance, while expansive, leaves much to be desired. Researchers do a poor job of tracking students from study beginning to end; few studies include the diverse populations found in public education; long-term academic benefits of exercise are largely unknown; few studies are replicated; there is very little qualitative research; and the quantitative studies do not control for important variables in monitoring academic performance, such as the quality of academic instruction. Looking ahead to future research, there is room for improvement in each of these areas.

Future studies should do a better job of data collection and reporting. Some attrition is expected, especially with larger populations, but the number of students lost in much of the current research is staggering. Furthermore, authors do not address this issue directly when reporting results and therefore do not appear objective. Future research should also do a better job of reporting student characteristics, especially important variables like SES and academic achievement. There is much room for rigorous, objective experimental studies that report student characteristics and the equivalence of groups used in the study.

Several study samples in the current literature were representative of the diversity found in public education. More often than not, however, they were of small homogeneous populations and findings were hard to generalize to other settings. Future research on diverse classrooms, or diverse schools, would add a great deal to the literature. This opportunity is made greater by the weakness of the research on comprehensive school wide physical activity programs. There have been few studies on these programs, and more research in diverse whole school settings is needed.

There is also room for more longitudinal studies on the benefits of exercise to academic performance. Only one study (Carlson et al., 2008) looked at more than three years of data, and findings were limited to correlations and not causal connections. Researching long-term benefits of exercise programs would help reveal not only whether benefits resulted from interventions, but also whether these benefits remained over time. Studies could also look at whether students maintained physical activity behaviors after formal interventions ended, and the results of these behavior patterns.

The body of literature would also benefit from more study replications. Only one study in this review (Pellegrini et al., 1995) conducted multiple replications of the same experiment. Other studies (Dwyer et al. 2001; Hill et al., 2011) alluded to replications but these were either not peer reviewed or were not pure replications. There is great value in researching a diversity of interventions, populations, measures, and outcomes; but results would have more weight if they were replicated in other experiments.

There is a noticeable lack of qualitative research regarding the effects of exercise on the academic performance of K-12 students. While this review was not exhaustive, qualitative studies were actively sought out and, for the most part, not found. The two such studies included in this review measured other outcomes besides academic performance; while stress, behavior, and self-esteem are valuable measures, there is a great deal of room for future qualitative studies to look at more direct measures of academic performance such as attention, concentration, and achievement. Case studies would provide evidence for connections between these measures and physical activity. Such studies would also allow individuals to offer their insights into why those connections exist in their particular cases. Interviews would also add to the body of research on this topic. Students, teachers, physical education instructors, brain specialists, and administrators could all contribute information regarding connections between types of physical activity and academic performance. These insights could be used to increase or decrease the amount of physical activity during the school day, alter types and durations of activity, or provide foci for further research. The richness, uniqueness, and detail of good qualitative studies would add a great deal to the discussion on the academic benefits of exercise.

Much of the research on physical activity and academic performance controls for baseline academic achievement of participants, and for physical activity during the study period. Conclusions are then drawn regarding the academic performance of students at the end of an exercise program. These methods miss an important piece of the puzzle: the academic instruction students receive concurrent with physical activity. This information is not important for studies of entire classrooms, or of a subset of students within a classroom. However, when studying an entire school population, part of a school population, or parts of classrooms, academic instruction becomes an important variable to consider. Data could reasonably be collected on time spent on particular subject areas, teacher experience and qualifications, time spent on homework, and other academic variables. Likewise, many studies did not control for physical activity that occurred outside of school hours, and those studies that did often relied on inconsistent and limited self-reporting. Future studies should account for these important variables as much as possible.

Lastly, the literature reviewed on connections between academic achievement and physical activity is limited to achievement in math, reading, and writing. These are valuable subjects, made more relevant by state and national standardized tests. However, studying the effect of physical activity on other content areas, such as social studies and science, would add to the research on this topic.

Conclusion

Chapter one of this review discussed the history of PE in the United States, in particular the link between emphasis on fitness and American involvement in wars as the military noted recruits' low fitness levels and policymakers viewed the issue as a threat to national security. Attention has, in the last couple decades, shifted to student achievement as measured by standardized tests. Chapter one also discussed the importance to educators of studying the link between physical activity and academic performance: policies on physical activity and academic performance should align with policymakers' goals according to research.

Chapter two summarized and analyzed the literature regarding the link between physical activity and academic performance, divided into five sections: fitness levels and academic performance, physical activity and academic performance, physical activity and academic achievement, physical exercise curricula changes and achievement test results, and the effects of physical activity on other areas affecting school performance. Findings in each of these sections were mixed. However, very few negative relationships were found between fitness and activity and measures of academic performance; in most cases, either a positive effect or relationship was found, or there were no effects or relationships.

Chapter three summarized the strengths and weaknesses of this collective research, which is diverse but often flawed. Chapter three also discussed the implications of these findings for classroom teachers, such as the benefits of conducting coordinated exercise lessons, taking short breaks during and between lessons, and using physical activity to decrease stress and increase self-esteem. Lastly, chapter three presented suggestions for future research, which included: more rigorous and objective studies, more qualitative studies, more research using diverse populations, and more consideration of academic variables such as classroom content and teacher experience.

There are many well-documented, universally accepted benefits of physical activity. However, this review is concerned with a less-documented topic: the connection between physical activity and academic performance. The research on this topic is varied in terms of interventions and connections studied, outcomes measured, and methods used. Most of the research has significant flaws; taken individually, each study would not hold much weight. However, taken as a whole, the research suggests that physical activity might help, and at the very least does not harm, academic performance. The research suggests, in particular, that students benefit from intentional and coordinated instruction, much like the intentional and coordinated instruction that guides effective classroom lessons. There is much room for further well-designed research in this area. For now, given the lack of detrimental effects to academic performance of physical activity interventions and the enormous costs that America's growing health crisis have presented to taxpayers, using schools as places for health education and physical activity makes sense. Administrators and the public should take note of these research findings, and teachers should use them to inform classroom instruction to the extent possible. Doing so may result in healthier, happier, and more knowledgeable classrooms.

References

Ahamed, Y., MacDonald, H., Reed, K., Naylor, P.-J., Liu-Ambrose, T., & McKay, H. (2007). School-based physical activity does not compromise children's academic performance. *Medicine & Science in Sports & Exercise, 39*(2), 371-376.

Budde, H., Voelcker-Rehage, C., Pietrabyk-Kendziorra, S., Ribeiro, P., & Tidow, G. (2008). Acute coordinative exercise improves attentional performance in adolescents. *Neuroscience Letters, 441*, 219-223.

Carlson, S. A., Fulton, J. E., Lee, S. M., Maynard, L. M., Brown, D. R., Kohl, H. W., III, & Dietz, W. H. (2008). PE and academic achievement in elementary school: Data from the early childhood longitudinal study. *American Journal of Public Health, 98*(4), 721-727.

Castelli, D. M., Hillman, C. J., Buck, S. M., & Erwin, H. E. (2007). Physical fitness and academic achievement in third- and fifth-grade students. *Journal of Sport & Exercise Psychology, 29*, 239-252.

Caterino, M. C., & Polak, E. D. (1999). Effects of two types of activity on the performance of second-, third-, and fourth-grade students on a test of concentration. *Perceptual and Motor Skills, 89*, 245-248.

Centers for Disease Control and Prevention. (2010). *The association between school based physical activity, including PE, and academic performance*. Atlanta, GA: Author.

Centers for Disease Control and Prevention. (2004). Participation in high school PE --- United States, 1991—2003. *MMWR, 53*(36) 844-847.

Centers for Disease Control and Prevention. (2011). *Childhood obesity facts*. Retrieved December 8, 2011 from http://www.cdc.gov/healthyyouth/obesity/facts.htm

Chomitz, V. R., Slining, M. M., McGowan, R. J., Mitchell, S. E., Dawson, G. F., & Hacker, K. A. (2009). Is there a relationship between physical fitness and academic achievement? Positive results from public school children in the northeastern United States. *Journal of School Health, 79*(1), 30-37.

Coe, D. P., Pivarnik, J. M., Womack, C. J., Reeves, M. J., & Malina, R. (2006). Effect of PE and activity levels on academic achievement in children. *Medicine & Science in Sports & Exercise, 38*(8), 1515-1519.

Corder, W. O. (1966). Effects of PE on the intellectual, physical, and social development of educable mentally retarded boys. *Exceptional Children, 32*(6), 357-364.

Crocker, J., & Park, L. E. (2004). The costly pursuit of self-esteem. *Psychological Bulletin, 130*(3), 392-414.

Dalleck, L. C., & Kravitz, L. (2002). The history of fitness. *IDEA Health and Fitness Source, 20*(2), 26-33.

DeBusk, M., & Hellison, D. (1989). Implementing a PE self-responsibility model for delinquency-prone youth. *Journal of Teaching in PE, 8*, 104-112.

Della Valle, J., Dunn, K., Dunn, R., Geisert, G., Sinatra, R., & Zenhausen, R. (1986). The effects of matching and mismatching on students' mobility preferences on recognition and memory tasks. *Journal of Educational Research, 79*(5), 267-272.

Dills, A. K., Morgan, H., & Rotthoff, K. (2011, April). *Recess, PE, and elementary school student outcomes*. Unpublished manuscript, Providence College, Providence, RI.

Dollman, J., Boshoff, K., & Dodd, G. (2006). The relationship between curriculum time for PE and literacy and numeracy standards in South Australian primary schools. *European PE Review, 12*(2), 151-163.

Donnelly, J. E., Greene, J. L., Gibson, C. A., Smith, B. K., Washburn, R. A., Sullivan, D. K., . . . Williams, S. L. (2009). Physical Activity Across the Curriculum (PAAC): A randomized controlled trial to promote physical activity and diminish overweight and obesity in elementary school children. *Preventive Medicine, 49*, 336-341.

DuBose, K. D., Mayo, M. S., Gibson, C. A., Green, J. L., Hill, J. O., Jacobsen, D. J., . . . Donnelly, J. E. (2008). Physical Activity Across the Curriculum (PAAC): Rationale and design. *Contemporary Clinical Trials, 29*(1), 83-93.

Dwyer, T., Sallis, J. F., Blizzard, L., Lazarus, R., & Dean, K. (2001). Relation of academic performance to physical activity and fitness in children. *Pediatric Exercise Science, 13*, 225-237.

Ericsson, I. (2008). Motor skills, attention and academic achievements. An intervention study in school years 1-3. *British Educational Research Journal, 34*(3), 301-313.

Finkelstein, E. A., Trogdon, J. G., Cohen, J. W., & Dietz, W. (2009). Annual medical spending attributable to obesity: Payer- and service-specific estimates. *Health Affairs, 28*(5), w822-w831.

Grissom, J. B. (2005). Physical fitness and academic achievement. *Journal of Exercise Physiology, 8*(1), 11-25.

Hellison, D. (1978). *Beyond balls and bats: Alienated (and other) youth in the gym*. Washington, DC: AAHPERD.

Hill, L. J. B., Williams, J. H. G., Aucott, L., Thomson, J., & Mon-Williams, M. (2011). How does exercise benefit performance on cognitive tests in primary-school pupils? *Developmental Medicine & Child Neurology, 53*(7), 630-635.

Hill, L., Williams, J. H. G., Aucott, L., Milne, J., Thomson, J., Greig, J., . . . Mon-Williams, M. (2010). Exercising attention within the classroom. *Developmental Medicine & Child Neurology, 52*(10), 929-934.

Hillman, C. H., Pontifex, M. B., Raine, L. B., Castelli, D. M., Hall, E. E., & Kramer, A. F. (2009). The effect of acute treadmill walking on cognitive control and academic achievement in preadolescent children. *Neuroscience, 159*, 1044-1054.

Hollar, D., Messiah, S. E., Lopez-Mitnik, G., Hollar, T. L., Almon, M., & Agatston, A. S. (2010). Effect of a two-year obesity prevention intervention on percentile changes in body mass index and academic performance in low-income elementary school children. *American Journal of Public Health, 110*(4), 646-653.

Knight, D., & Rizzuto, T. (1993). Relations for children in grades, 2, 3, and 4 between balance skills and academic achievement. *Perceptual and Motor Skills, 76*, 1296-1298.

McCormick, C. C., Schnobrich, J. N., Footlik, S. W., & Poetker, B. (1968). Improvement in reading achievement through perceptual-motor training. *The Research Quarterly, 39*(3), 627-633.

McNaughten, D., & Gabbard, C. (1993). Physical exertion and immediate mental performance of sixth-grade children. *Perceptual and Motor Skills, 77*, 1155-1159.

Oliver, J. N. (1958). The effect of physical conditioning exercises and activities on the mental characteristics of educationally sub-normal boys. *British Journal of Educational Psychology, 28*, 155-165.

Pellegrini, A. D., Huberty, P. D., & Jones, I. (1995). The effects of recess timing on children's playground and classroom behaviors. *American Educational Research Journal, 32*(4), 845-864.

Phelps, R. P. (2011). The effect of testing on achievement: Meta-analyses and research summary, 1910-2010. *Nonpartisan Education Review, 7*(2), 1-25.

Pick, A., Frankel, D., & Hess, V. (1975). Children's attention: The development of selectivity. In E. M. Hetherington (Ed.), *Review of child development research* (vol. 5, pp. 325-383). Chicago: University of Chicago Press.

Pollatschek, J. L., & Hagan, F. J. O. (1989). An investigation of the psycho-physical influences of a quality daily PE program. *Health Education Research, 4*(3), 341-35.

Popham, W. J. (1999). Why standardized test scores don't measure educational quality. *Educational Leadership, 56*(6), 8-15.

Ravitch, D. (2010). *The death and life of the great American school system: How testing and choice are undermining education*. New York: Basic Books.

Raviv, S., & Low, M. (1990). Influence of physical activity on concentration among junior high-school students. *Perceptual and Motor Skills, 70*(1), 67-74.

Sallis, J. F., McKenzie, T. L., Kolody, B., Lewis, M., Marshall, S., & Rosengard, P. (1999). Effects of health-related PE on academic achievement: Project SPARK. *Research Quarterly for Exercise and Sport, 70*(2), 127-134.

Seldon, T. M., & Sing, M. (2008). The distribution of public spending for health care in the United States, 2002. *Health Affairs, 27*(5), 349-359.

Shepard, R. J., (1996). Habitual physical activity and academic performance. *Nutrition Reviews, 54*(4), S32-S36.

Storey, H. C., Pearce, J., Ashfield-Watt, P. A. L., Wood, L., Baines, E., & Nelson, M. (2011). A randomized controlled trial of the effect of school food and dining room modifications on classroom behaviour in secondary school children. *European Journal of Clinical Nutrition, 65*, 32-38.

Tergerson, J. L., & King, K. A. (2002). Do perceived cues, benefits, and barriers to physical activity differ between male and female adolescents? *Journal of School Health, 72*(9), 374-380.

Tremarche, P. V., Robinson, E. M., & Graham, L. B. (2007). PE and its effect on elementary testing results. *Physical Educator, 64*(2), 58-64.

Tremblay, M. S., Inman, J. W., & Willms, J. D. (2000). The relationship between physical activity, self-esteem, and academic achievement in 12-year-old children. *Pediatric Exercise Science, 12*, 312-323.

Tuckman, B. W., & Hinkle, J. S. (1986). An experimental study of the physical and psychological effects of aerobic exercise on schoolchildren. *Health Psychology, 5*(3), 197-207.

U.S. Department of Health and Human Services. (2008, June). *Physical Activity Guidelines Advisory Committee report*. Washington, DC: Author.

Vincent, D., & Cresswell, M. (1976). *Reading tests in the classroom*. Slough, England: National Foundation for Educational Research.

Vinciullo, F. M., & Bradley, B. J. (2009). A correlational study of the relationship between a coordinated school health program and school achievement: A case for school health. *The Journal of School Nursing, 25*(6), 453-465.

Welk, G. J., Morrow, J. R., & Falls, H. B. (2002). *Fitnessgram reference guide*. Dallas, TX: The Cooper Institute.

Weston, A. T., Petosa, R., & Pate, R. R. (1997). Validation of an instrument for measurement of physical activity in youth. *Medicine & Science in Sports & Exercise, 29*, 138-143.

Zervas, Y, Danis, A., & Klissouras, V. (1991). Influence of physical exertion on mental performance with reference to training. *Perceptual and Motor Skills, 72*, 1215-1221.