

Rowan University

Rowan Digital Works

---

Theses and Dissertations

---

5-12-2004

## The effects of Success for All as a whole school reform on the GEPA scores in a particular Abbott District

Lisa R. Labbree  
*Rowan University*

Follow this and additional works at: <https://rdw.rowan.edu/etd>



Part of the [Special Education and Teaching Commons](#)

---

### Recommended Citation

Labbree, Lisa R., "The effects of Success for All as a whole school reform on the GEPA scores in a particular Abbott District" (2004). *Theses and Dissertations*. 1183.  
<https://rdw.rowan.edu/etd/1183>

This Thesis is brought to you for free and open access by Rowan Digital Works. It has been accepted for inclusion in Theses and Dissertations by an authorized administrator of Rowan Digital Works. For more information, please contact [graduateresearch@rowan.edu](mailto:graduateresearch@rowan.edu).

THE EFFECTS OF SUCCESS FOR ALL AS A WHOLE SCHOOL REFORM ON THE  
GEPA SCORES IN A PARTICULAR ABBOTT DISTRICT

By  
Lisa R. Labbree

A Thesis

Submitted in partial fulfillment of the requirements of the  
Master of Arts Degree  
Of  
The Graduate School  
At  
Rowan University  
May, 2004

Approved by \_\_\_\_\_  
Professor

Date Approved 5/12/04

## ABSTRACT

Lisa R. Labbree

THE EFFECTS OF SUCCESS FOR ALL AS A WHOLE SCHOOL REFORM ON THE  
GEPA SCORES OF A PARTICULAR ABBOTT DISTRICT

2003/2004

Dr. Stephen Crites

Master of Arts in Special Education

The purpose of this study was to determine what impact Success For All (SFA) had on the state standardized test scores of a school district that was identified as being low income and having poor test scores. Data from the literacy/language arts portion of the Grade Eight Proficiency Assessment were compared for the years 2001 and 2003. Students who took the test in 2001 were not exposed to SFA, whereas the 2003 students had three years of SFA. The test scores were analyzed by comparing percentage passed, mean scores, and mean differences. Results indicated that the test scores for special education students and the lowest performing 25% improved slightly, while the regular education population scores decreased.

## Acknowledgments

Everyone who has had to deal with me this academic year should be acknowledged. This goes especially for my family who has had to live with the trials and tribulations of this thesis. My husband, John Banning, has willingly given up his own time to assist me when my computer inadequacies got the best of me. His loving support has helped me through this process. As always, I appreciate his patience and perseverance. My children, Cole and Shane, have given up much computer time and have admirably dealt with my impatience and moodiness. I also thank them for their support and patience. My mother, Jean Labbree, has given up her time to help with the proofreading of this, at times tedious, paper. I have always appreciated the emotional support of both of my parents throughout the years.

I also must mention the support and patience of my friends and other family members. This past school year, I have neglected them and they have stood by me with understanding. I hope I can somehow repay the favor.

Of course, my thesis advisor, Dr. Steven Crites, must be acknowledged for his time and effort. He always made me believe that my paper was worthwhile, which helped me stay motivated. I thank him for his suggestions, edits, and encouragement.

## Table of Contents

|      |                          |                 |       |
|------|--------------------------|-----------------|-------|
| I.   | Introduction             |                 | p.1   |
| II.  | Review of the Literature |                 | p. 7  |
| III. | Methodology              |                 | p. 43 |
| IV.  | Results                  |                 | p. 48 |
| V.   | Discussion               |                 | p. 53 |
| VI.  | References               |                 | p. 61 |
| VII. | Appendix A               | IRB Letter      | p. 70 |
|      | B                        | Approval Letter | p. 71 |

## Chapter 1

### Introduction

There is much discussion in education today about the inequities that exist in the nation's school systems. The reality of unequal educational opportunities has been exhibited in many ways and has always been a problem for the public school system (Rossmiller, 1994). Although there seems to be some agreement that this inequity exists in society, there is much discussion on the reasons or solutions for this educational dilemma. Unequal and inconsistent funding policies (Goertz, 1994), societal poverty and racial discrimination (Karp, 1997; Lowe, 1997; Clune, 1994), and classroom instructional practices (Clune, 1994) are some of the reasons given for this particular problem in our schools. There are numerous consequences due to this type of inequity. Unequal educational opportunity is viewed by many as responsible for high dropout rates, inadequate and antiquated buildings, increased student absenteeism, and low performance on standardized tests.

There have been many attempts over the years to try to enhance educational opportunity in schools that exist in poor communities that have low student achievement. Many of the education reforms have resulted from the concept of unequal funding (Karp, 1997). Federal and state legislation have had an influence in this area. As early as 1965, the federal government allocated extra funds to poor districts through Title 1/Chapter 1 of the Elementary and Secondary Education Act (ESEA). Throughout the years, this program has increased funding and has made it easier for schools to be eligible for the funds. In 1994, districts were able to receive money if they could show 50% or more of

the students were below the poverty level as shown by recipients of free or reduced lunch. In 1997, Congress allocated an additional \$145 million for low-performing schools, mostly Title 1 schools, as an incentive to use research-based instruction (Herman, 1999).

Although not its intention, Title 1 became a type of “pull-out” program for disadvantaged students. Low performing students were identified and taken out of regular classes for supplemental instruction. As stated by D’Agostino, Borman, Hedges, and Wong (1998), “...long-standing legal provisions and implementation mechanisms designed to target the delivery of supplementary educational services to students placed at risk have continued to stress the structural separation of categorical services from local policy and practice” (p.402). The focus was on the individual student and his/her lack of basic skills. However, some research has indicated that these separate pullout classes do not give the students the same opportunity to the curriculum of the regular classroom. As a result, the achievement level of the Title 1 students has not improved. This type of program did not have the intended effect of raising performance levels (Kantor, 1997; D’Agostino et al, 1998).

The idea of reforming the entire composition of the school, and not just the low-performing students, can be traced back many decades. From Bertand Russell and John Dewey in the early 1900’s to John Holt and James Comer in the 1960’s, educational visions and movements have existed to help improve our schools. The idea of whole school reform is “newer still” (Traub, 1999, p.7)

In the 1980s, there were many individuals who devised and implemented whole school designs. These programs looked at the entire school community as well as the

community services available outside the school. (Traub, 1999). In this type of educational change known as whole school reform (WSR), “ educational improvement is envisioned as encompassing changes in all the critical facets of a school’s environment in a coordinated and systematic manner” (Walker& Gutmore, 2000, p.2). Examples of these programs include Accelerated Schools, Core Knowledge, Comer Project, Success For All, and Edison School.

In the 1990s, this idea of servicing the entire school, and not just the economically disadvantaged, flourished with the institution of the New American Schools Development Corporation. This program is still in existence today under the name New American Schools (NAS). This program was designed to give financial support to developers of educational designs that incorporated the entire school (Traub, 1999). Additionally, the Comprehensive School Reform Demonstration (CSRDP) program (P.L. 105-78) was developed in 1998 by the U.S. Department of Education to give funding to schools that use scientifically-based school reforms (Doherty, 2000). This law stipulates nine requirements that a reform must meet for the school to receive funding. They are as follows: (1) effective, research-based methods and strategies; (2) comprehensive design with aligned components; (3) professional development; (4) measurable goals and benchmarks; (5) support within the school; (6) parental and community involvement; (7) external technical support and assistance; (8) evaluation strategies; and (9) coordination of resources. (Traub, 1999; Doherty, 2000). In 1994, Title 1 money was available to high poverty districts to use for the entire school population (Fashola & Slavin, 1998).

Also in 1994, the US Congress passed Goals 2000 legislation in an attempt to provide minimum academic standards as a means to help the students who come from



high poverty districts. However, in order to get the legislation passed, the bill was diluted and each state set the standards and assessment of the standards for themselves. Each state, then, could determine what was to be done to achieve educational equity (Orfield, 1994). Many states incorporated standardized testing that matched their academic standards as an attempt to analyze whether the students were achieving these standards.

President George W. Bush signed the most recent legislation in January 2001 entitled No Child Left Behind Act (NCLB), P.L. 107-110. This was a reauthorization of the Elementary and Secondary Education Act of 1965. Title 1, Part A of this law provides extra resources to high-poverty schools, as did the original legislation. Among other changes, this legislation gives more flexibility with Title 1 monies, changing the criteria for funding to 40% free or reduced lunch students (U. S. Department of Education [USDOE], n.d.a). Also tied in with Title 1 funding is the mandate of increased accountability of school districts and a provision that schools that are failing show adequate yearly progress (AYP) through an increase in standardized test scores. NCLB act also mandates the use of research-based instruction in schools that are not showing academic progress (USDOE, n.d.c).

#### Statement of problem

New Jersey has identified, and is using, many models of whole school reform (NJDOE, 2000a). There is some question as to whether all these models are scientifically-based educational programs (Borman, Hewes, Overman, & Brown, 2002). Since this idea of the state being involved in mandating whole school reform is relatively new, there is a need for studies to show the effectiveness of this approach in gaining educational equity. One of the criteria at the state and national level for achieving

equality is the closing of the achievement gap through testing of all students (Herman, 1999; Springfield, 2000). It is imperative to determine whether the state-approved reform models are actually improving education as seen in test scores. These models are comprehensive but expensive (Herman, 1999). Studies so far have been inconclusive on even the most studied reform models; e.g., Comer, Success for All, Coalition of Essential Schools, Direct Instruction (Borman, Hewes, & Brown, 2002; Traub, 1998; Springfield, 2000). With federal and state governments operating under fiscal constraints, funding for education needs to be cost-effective. This study will investigate the impact of a particular whole school reform model, Success for All (SFA), on the language art literacy portion of the state test given to eighth graders, the Grade Eight Proficiency Assessment (GEPA), in a low-income school district in southern New Jersey.

#### Justification for study

There are several studies on the impact of SFA on test scores. However, they are controversial in nature. Much research is done by people associated with the creator of SFA, and many critics argue this taints the result (Greenberg & Walberg, 1999). In addition, most of the studies are done at the elementary level. This study will extend the knowledge base into the middle school level.

There is also the question of what happens to classified students, those students who are identified as having specific learning problems, when whole school reform takes them out of the resource center and away from special education teachers. This research project will also analyze the students who would have been in the resource center but received SFA instruction instead. This information could impact the way classified students are taught and add to the knowledge base on inclusion.

Another aspect that makes this study different is the control group. Most studies involving whole school reform have a control school that they try to match to the school receiving the school reform. This study will compare two sets of students in the same school. Using students from the same school as the subjects could control any intervening variables that may exist when comparing different schools.

Also, this study is unique due to the composition of the sample. Most special needs districts are comprised of a large percentage of minorities. This complicates the research by adding another intervening variable. The sample and control group for this study is homogeneous; the population is 98% white. In this way, the research can center on the low-income of the district without race becoming a factor.

#### Research Questions

1. Did the students as a whole group who received the whole school reform model do significantly better on their literacy/language arts GEPA scores than those students who were not exposed to this program?
2. Did the regular education students who received the whole school reform model do significantly better on their literacy/language arts GEPA scores than those who were not exposed to this program?
3. Did the classified students who received the whole school reform model do significantly better on their literacy/language arts GEPA scores than those who were not exposed to this program?
4. Did the lowest performing group of students do significantly better on the literacy/language arts scores than those that were not exposed to this program?

## Chapter 2

### Review of the literature

States are also involved in the trend to end educational inequity through increased spending and school accountability. In addition to states such as Texas and Kentucky, New Jersey is in the forefront of this type of educational and financial reform (Clune, 1994; Slavin, 1994). There is a long history in New Jersey of legislation involving unequal opportunities being attributed to inadequate funding of poverty districts because of the state's reliance on property taxes to fund schools. In 1973 (Education Law Center [ELC], n.d.b), *Robinson vs. Cahill* (62 NJ 473) determined that New Jersey's funding of schools was unconstitutional. This resulted in a funding formula through the Public School Educational Act of 1975 that used a minimum funding on a per pupil basis. However, this new formula did not change the disparity of spending for wealthy and poor districts. This act was challenged by four low-income districts in 1988 under the auspices of the Education Law Center and found also to be unconstitutional; this case is known as *Abbott vs. Burke*.

In *Abbott vs. Burke II*, the State Supreme Court mandated the state to provide additional services to assist the students in high-poverty areas. The court identified 28 districts as special needs districts (ELC, n.d.c); in 1998, this number was changed to 30 (ELC, n.d.b). The New Jersey government in response came up with the Quality Education Act that set up a funding formula that increased aid to the poor districts and restricted aid to the wealthier districts. This, too, was found to be unconstitutional since it didn't ensure equality of funds or provide supplemental services to the poor districts

(Abbott III). The response of the legislature at this point was the development of standards in seven core content areas, known now as the Core Content Curriculum Standards, and a minimum per pupil expenditure of \$6,720. The State Supreme Court ruled this response unconstitutional as well. The court ruled the per pupil expenditure arbitrary and lacking in any real knowledge of the needs of the school districts; this case is known as Abbott IV (Walker & Gutmore, 2000).

As a result of Abbott IV, a study was required by the State Department of Education to determine the special needs of the identified districts. In September 1997, the state complied with the ruling by earmarking \$246 million dollars that resulted in equal spending between wealthy and poor districts for the first time (ELC, n.d.c). Abbott V in 1998 instituted the mandate of Whole School Reform (WSR) programs being used in an attempt to achieve educational equity as evidenced by a shrinking of the achievement gap (New Jersey Department of Education [NJDOE], 2000). Subsequent cases involving Abbott vs. Burke have revolved around additional schools seeking the special needs status and districts trying to get monies for mandated supplemental programs, such as preschool (ELC, n.d.c).

According to the New Jersey Department of Education (2000), the current Abbott remedies can be divided into four broad substantive categories: standards-based reform, early childhood education, social and health services and other services and facilities improvement. The remedies incorporate the elements of equity, efficiency and excellence and attempt to redistribute educational funding within a prescribed formula. Within the standards-based reform context, whole school reform using designated models is required by all schools in three year phases

and a governance model of site-based budgeting and decision making is further required to guide the implementation of the designated reform model (p.4).

Due to the fact that WSR was instituted to increase achievement levels in low-income districts, the effectiveness of the various models must be assessed. However, there are many difficulties in evaluating them (Borman, Hewes, Overman, & Brown, 2002; Springfield, 2000; Herman, 1999; Traub, 1999; Greenberg & Walberg, 1998). As Traub (1999) points out, since there is a difference in philosophy of scientific theory, academic problems, and best practices for the various models, comparing the programs is problematic. He also accentuates the fact that the terms involved in evaluation are not consistent; for instance, the use of the word “standards” is controversial since the individual schools many times decide upon the interpretation.

Greenberg and Walberg (1998) agree that this presents problems in researching programs. In their review of evaluation literature, they have identified additional biases that could influence evaluation. They comment on the fact that very few studies are done by independent, or third party, researchers. Other researchers have come to the same conclusions (Herman, 1999; Springfield, 2000; Borman et al., 2002).

Additionally, many of the studies, as well as the programs, are funded by governmental agencies that have a vested interest in their success; for example, the federal government and its involvement with Title 1 (Greenberg & Walberg, 1998). Moreover, Greenberg and Walberg have indicated as many as 20 factors that could lead to bias in their outcomes; these include, but are not limited to, biases in design and the choice of the design, biases in reporting the data and the implementation of the program,

biases in the instrument and testing, and biases in the selection process for program review.

Borman et al. (2002) also identified several problem areas within the field of program evaluation. In their meta-analysis of WSR and its effect on student achievement, they discuss the problem of comparing and contrasting studies when there are differences in who reported the outcomes, the methodology used, the difference in school settings, the characteristics of the program (cost, level of support), and how the program assesses its own success (various types of test scores).

There are currently many guides to evaluate the programs that are within the scope of WSR (Borman et al., 2002; Herman, 1999; NWREL, 1998; Slavin & Fashola, 1998; Wang, Haertel, & Walberg, 1997; AFT, 1998; Traub, 1997). The educational community generally perceives these reports as reviews of current reforms for the purpose of explaining the program, evaluating the costs, and presenting the outcome data (Herman, 1999). Programs included in these guides usually have to adhere to the components that the U.S. Department of Education developed as part of the Comprehensive School Reform Demonstration (CSRD); they have numerous implementation sites, have been subject to empirical study, and require developer input (Borman et al.). Programs that are frequently included are Accelerated Schools, America's Choice, Coalition of Essential Schools, Core Knowledge, Direct Instruction, Edison Project, High Schools That Work, Modern Red School House, New American Schools, School Development Program, and Success for All (Borman et al.; Herman; NWREL, 1998; Slavin & Fashola, 1998; Wang, Haertel, & Walberg, 1997; AFT, 1998; Traub, 1997).

*An Educator's Guide to Schoolwide Reform* (Herman, 1999) is one of the more comprehensive guides to reform program review (Springfield, 2000; Slavin & Madden, 2001b). It is viewed as one of the few guides that is not tainted with bias (Springfield, 2000; Traub, 1999). This guide "was funded by five of the most important independent groups in U.S. education" (Springfield, p.265). This enables an objectivity that is lacking in other reviews (Traub, 1999). Another way that this report is superior to some of the others is that it only includes programs that were studied using quasi-experimental design (Herman, 1999; Springfield). Of the 24 programs that this group evaluated, only two programs showed strong evidence of impact on student achievement, Success for All and Direct Instruction (Herman; Springfield).

Traub (1999) has cited some limitations that occur in this review of reform programs. He questions what the Herman (1999) study was actually considering as significant. It is Traub's contention that this particular guide gave higher marks to programs that had many studies despite the presence of other programs that seemed to have an empirical basis; in this way, he states, older programs have an advantage. Additionally, the study treats all positive effect sizes equally, whether they are large or small.

Another guide that evaluates comprehensive reform models is *Show Me the Evidence!* (Slavin & Fashola, 1998). This guide also used the requirement of quasi-experimental design studies as a criterion for admission. However, the guide did not adhere to its own plan. In reality, many programs were evaluated that did not have the same methodological basis (Springfield, 2000). More importantly, Robert Slavin is the founder of Success for All, one of the programs that is evaluated in the study. This, of



course, presents a conflict of interest that could affect the outcome of the evaluation (Jones, Gottfredson, & Gottfredson, 1997); Pogrow, 2000a).

Another guide that is prominent in this area is entitled *What Do we Know?* that was done by Wang, Haertel, and Walberg (1997). The U.S. Department of Education's Office of Educational Research and Improvement (OERI) funded the study. This guide of 12 programs was updated recently on the Internet. The updated version is more inclusive than the original. Each program was evaluated using 80 different criterion. There are two significant problems with this research. First, the guide does not include any achievement data; the authors have indicated that they chose not to include this information due to the insignificant results. However, other researchers have found significant data in this area. Secondly, there may be a conflict of interest in that Wang is the developer of Community for Learning program. Interestingly enough, this program met most of the criterion identified by the researchers (Springfield, 2000).

A fourth guide was compiled by Northwest Regional Educational Laboratory (NWREL) entitled *Catalogue of School Reform Models* (1998) and was also funded by OERI. This guide reviews 33 WSR models in a mostly descriptive way; its evaluation includes all the studies done on a program. The fact that all studies, regardless of their scientific value, are treated equally presents a problem in interpreting the information given.

Three additional guides should be mentioned. James Traub (1999) evaluated ten programs under the auspices of The Thomas B. Fordham Foundation. Besides including an essay on the educational philosophy of each program and an example of actual implementation of the program, he charts, using a five-point system, the effectiveness of

the program on student achievement and the amount of program support given to the school. Only Success for All received high marks on both accounts. He does conclude, however, that “we cannot yet definitely say that some designs work better than others do, or that any one design is successful with all children in all situations” (p. 10).

Calling other guides outdated, Borman et al. (2002) evaluated, using meta-analysis, 29 of what they considered the most widely implemented school reform models. To be included in the study, a program had to be studied at least once and had to be implemented in at least ten schools. This research did not include all studies involving the particular program; it concentrated on the scientifically based research that focused on student achievement. The program guides indicated above were part of this meta-analysis. There were several interesting findings of this study. The researchers state that there are limitations to the quantity and quality of studies supporting achievement effects; the effects of WSR programs are statistically significant and are greater than non-WSR interventions; there is a lack of continuity of reform characteristics that these programs share; developers’ evaluations show a larger effect size than independent evaluations; the three programs that have been researched in a variety of ways have the largest effect size; and the longer a school has a WSR model, the higher the achievement levels. Based on the researchers’ criteria, the three programs that had the best outcomes are Direct Instruction, the School Development Program, and Success for All.

The American Federation of Teachers [AFT] (1998) has also been involved in evaluating WSR programs. Through the AFT Task Force on Improving Low-Performing Schools, a guide was assembled that identified six programs that had scientific evidence of raising student achievement. In order to be included in this particular guide, a program

had to show an effect size of at least .25, which they state is educationally significant. Also, the program had to have had third party evaluations at multiple sites. Furthermore, they had to have comparison data to control groups or some type of standardized test score. Included in this compilation were Success for All, Direct Instruction, Core Knowledge, and Early Steps. Interestingly enough, two other school reform programs that did not have an academic component showed evidence that the achievement scores were raised; these programs were School Development Program and Consistency Management and Cooperative Discipline.

SFA has been one of the most studied WSR programs (Borman & Hewes, 2002; Slavin, Madden, Dolan, & Wasik, 1996). Currently there are 1,500 schools in 48 states that are using this comprehensive, WSR model. According to the non-profit Success for All Foundation (2003), this program has been evaluated in 47 experimental-control group studies. Approximately 70% of SFA students are African American or Latino (Slavin & Madden, 2001a). Comparisons have been made using individual reading scales and state administered tests (Slavin & Madden, 2001b).

This WSR grew out of research being conducted by Robert Slavin and Nancy Madden on various techniques that could best educate at-risk students. This research dates back to the 1970s when the pair was researching cooperative learning. In 1987 the Baltimore school system elicited the help of Slavin and Madden who were working at the Center for Social Organization of Schools at John Hopkins University. Baltimore's superintendent was trying to locate a program that would enhance the learning potential of at risk students.

“At-risk” was defined by the U.S. Department of Education through the Research, Development, Dissemination and Improvement Act of 1994 as someone who is placed at greater risk of low academic achievement due to limited English speaking, poverty, race, geographic location, or economic disadvantage (National Institute on the Education of At-risk Students [NIEARS], 1998).

In 1997 SFA was identified in the CSRD as a presumptive model (Slavin & Madden, 2000). In 1998 SFA was recommended as the preferred WSR model in the Abbott legislation (Slavin & Madden, 1999). Additionally, extra funding of \$25,000 was guaranteed to any Abbott school choosing this model (Crosbie, 2000).

SFA “is a program designed to comprehensively restructure elementary schools serving many children placed at risk of school failure. It emphasizes prevention, early intervention, use of innovative reading, writing and language arts curriculum, and extensive professional development to help schools start children with success and then build on that foundation throughout the elementary grades” (Slavin, Madden, Dolan, & Wasik, 1996, p.198).

There are several elements that are crucial to the SFA program. One of the most important is the reading component. The philosophy of this component stems from best practices on reading instruction using a schoolwide curriculum. Students are grouped according to reading level, not grade level. The homogeneous reading groups meet during the same time of day for 90 minutes. Much of the reading/language arts instruction involves cooperative learning. Reading for twenty minutes a night is assigned for homework. Students who are having difficulty in reading may get an additional twenty minutes of instruction by tutors. These reading tutors are usually certified

teachers. Slavin (Slavin, Madden, Dolan, & Wasik, 1996) asserts that this is especially important for the achievement of the lowest 25% of the students; their success is correlated with the number and quality of the tutors. He recommends a 1:5 ratio of tutor to student. A major goal of the SFA program is to have all students reading at grade level by third grade (Madden, 1992); hence, success for all.

Another essential part of the program is the assessments given every eight weeks. Assessment is done to appraise the progress, or lack of progress, of individual students. If a student is having trouble, some type of intervention is made; this could entail a change in grouping, a tutor assignment, or some type of emotional/behavioral assistance.

Involving the family is an added feature of SFA. A Family Support Team is designed to give support and education to the family and the child. This team consists of a social worker, counselor, SFA facilitator, and vice principal. Issues that are commonly addressed by this team are attendance, health, and emotional/behavioral. Students may be referred by the reading/language arts teacher or identified by performance on the eight-week assessment. According to Slavin (Slavin, Madden, Dolan, & Wasik, 1996), the component of family support is especially vital to the philosophy of helping all students and has facilitated an increase in attendance for the lower performing students.

Professional development and continued support from SFA are other elements that are essential to the program. SFA is initially instituted by a three day in-service to train teachers/tutors in the instructional techniques and in the use of the teacher manuals. Tutors receive two additional days of in-service in strategies and assessment. Throughout the year, there are other in-services and/or informal workshops to assist in the implementation of the program.

The SFA facilitator oversees the SFA program in each school. Someone who is already on staff at the school usually fills this position. The responsibilities include planning and scheduling; assisting the teachers with curriculum and classroom management; managing the eight week assessments; and coordinating activities of the Family Support Team (Slavin & Madden, 1996; Slavin & Madden, 2001; National Institute on the Education of At-Risk Students [NIEARS], 1998).

Professional development and continued support from SFA are other elements that are essential to the program. SFA is initially instituted by a three day in-service to train teachers/tutors in the instructional techniques and in the use of the teacher manuals. Tutors receive two additional days of in-service in strategies and assessment. Throughout the year, there are other in-services and/or informal workshops to assist in the implementation of the program (NIEARS, 1998).

Determining the cost of this program is somewhat problematic. Several studies have determined that this particular model is the most expensive (Borman & Hewes, 2002; Herman, 1999; King, 1994). Herman estimated the first year costs of the program to be close to \$270,000 for personnel, materials, and training while King's estimate was even higher, up to \$646,500. Slavin (Slavin & Madden, 2001b), however, has taken exception to these estimates. He has argued that schools have been able to be creative in using teachers and staff that are already available to them, so that the money needed for staffing is misleading. He also has argued that this isn't additional money needed because schools have reallocated funding from other sources, such as Title 1 and special education. Many times, special education teachers are taken out of the resource center to teach reading/language arts to the lowest performing students (Slavin & Madden).

SFA estimates their cost to be \$60,000 to start (The NIEARS, 1998) with initial cost of materials estimated to be \$20,000 (Borman & Hewes, 2002). Professional development costs about \$800 per SFA personnel per day.

Another important quality of SFA is that in order for a school to initiate the program, the developers insist on a consensus of the staff. After learning about the program, the staff must vote and achieve at least an 80% agreement among them. If this type of consensus is not met, the program will not be instituted (Slavin & Madden, 2001b).

Proponents of SFA argue that schools that implement the program are more cost-effective (Borman & Hewes, 2002). Primarily, savings come from the SFA outcomes of reducing retentions and special education placements (Slavin & Madden, 2000). It is also contended that the cost of materials is counteracted by the fact that educational materials would have been ordered regardless, so that this is not an additional cost (Slavin, Madden, Dolan, & Wasik, 1996).

According to Slavin and Madden (2001b), this program was designed to be researched and replicated. In the 16 years since its inception, there have been numerous studies on this program involving different school districts in several states ( Slavin & Madden, 2003a). Many have used the longitudinal design model; others have used survey and observation. Even though studies on SFA included individual and group testing, Slavin and Madden (1997) stated that the individually administered assessments using Woodcock Language Proficiency Battery and the Durrell Analysis of Reading Difficulty are more accurate than group standardized tests and are more sensitive to actual increase in reading scores. Many of SFA's testing is done orally (Slavin & Madden, 2001a).

Ross and Smith (1996) contend that this type of measurement is more sensitive to reading differences so it is a better indicator of reading improvement than other types of testing.

The first school to use SFA was matched with a similar school and followed over several years (Slavin, Madden, Karweit, Livermon, & Dolan, 1990). The reform was instituted in pre-kindergarten to third grade in a Baltimore, Maryland school that had a predominately black population and had 76% of students qualifying for free/reduced lunch. A nearby school that matched percentage of free lunch, achievement level, and ethnicity was used as a control group. SFA has used this type of quasi-experimental design for subsequent research (Slavin & Madden, 2003a). The researchers match students through existing test scores and posttests of individually administered tests. Several types of tests were used for individual and group scores. For pre-K and first grade, the tests included the Test of Language Development (TOLD) and the Merrill Language Screening Test for language. For first through third grade, the Woodcock Language Proficiency Battery, the Durrell Analysis of Reading Difficulty, and California Achievement Tests (CAT) were used for assessing reading scores. The CAT test is a group administered reading test.

Results were evaluated after the first year of implementation. Data were discussed in terms of effect sizes and analyzed using covariances. Large effects were seen on the individual tests, but no effects were seen on the CAT tests. Slavin et al. (1990) contend that this was due to the control group's curriculum being closely aligned to the standardized test.

Third grade had the largest difference between the control group and the SFA school as far as effect size on the individually administered tests, with an average of ES



+0.95. There also was a large difference in effect size for the lowest performing 25% of the third grade. In addition, retention and special education referrals were reduced. This school district, including this school, was studied for many years.

Other studies were done in the Baltimore school district. Madden, Slavin, Karweit, Dolan, and Wasik (1993) studied five SFA schools over a period of three years; the schools were matched with control schools. The researchers were interested in comparing individual reading test scores for grades one through three. The tests used were the Woodcock Language Proficiency Battery and the Durrell Analysis of Reading Difficulty. Only those students who were in the schools throughout the three-year period were included. In general, the data indicated that the SFA students outperformed the control students. There were positive significant effects on the multivariate analyses for most students, with the second grade at one school being the exception. Moreover, there were positive effect sizes for all the reading measures, ranging from +0.82 to +1.00. There were more control students reading at least one year below grade level (50% as compared to 20% SFA) at the conclusion of this research. Grade retentions were reduced and attendance improved in the SFA schools. The researchers attributed these improvements to the family support team.

Richard Venesky (1998) was hired as an independent researcher to review the research done by Madden et al. (1993). He questioned the methodology used by the SFA researchers as well as their interpretation of the results. He contended that using research by developers of the program might not be the best idea. In addition, he asked whether control group comparisons really validated a program's success. He pointed out that SFA fifth graders were 2.4 years behind the national norm in reading. He also pointed out that

the SFA schools had additional funding given to them, up to \$400,000 (Pogrow, 2000a), so that they had an unfair advantage over the control schools.

Venesky (1998) also took offense to the comparison for the individually administered tests to national norms. He stated that although they have some norming value, they should not be used in the way SFA uses them. Other researchers concur with Venesky. Jones, Gottfredson, & Gottfredson (1997) commented that the individually administered tests do not correspond to the way schools usually evaluate their students. They contend that there is an inherent problem in this type of analysis in finding a control match and comparing the results afterward.

In protest, Slavin and Madden (2000) argued that the SFA students did better than their control counterparts, scoring a full grade ahead on the individually administered tests and slightly below a full grade on the CTBS, Comprehensive Test of Basic Skills. They also justify their use of the Woodcock, Durrell, and Gray tests and reiterate that the SFA schools are originally the worst performing schools in the district.

The Department of Research and Evaluation (Ruffini, Feldman, Edirisooriya, Howe, & Borders, 1992) conducted several analyses on the Baltimore City Public Schools' student database for the years 1988-1991. Although this assessment qualifies their results with the point that they could not determine if the Baltimore schools successfully implemented the SFA program as designed, they were not optimistic about the results. They commented that the majority of the SFA students were not reading on grade level after years of having SFA. They questioned why if reading on grade level by third grade was a primary purpose, the methodology did not test this. Instead it compared SFA students with a control group.

In addition, Ruffini et al. (1991) expressed concern that the SFA schools that were supposed to have the top of the line program had the lowest scores in the district. There were 28 SFA schools in the district; 19 met their goals, but 9 did not. Ruffini et al. concluded that they did not think the program was meeting the needs of the students.

Another problem Ruffini et al. (1991) had was with the claim that SFA reduced special education placement; they concluded that SFA students who went to different schools were referred for special education at the same rate as the control schools. According to SFA, not referring a student for special education is a program policy, not a result of the program (Slavin & Madden, 2001b). Jones et al. (1997) stated that SFA researchers should not use the statistics of special education placement and retentions as evidence of the success of the program due to the fact that not referring and not retaining are intrinsic to the policies of SFA.

In fact, Slavin (1996) referred to the policy of not placing students in special education as “neverstreaming.” He theorized that if students were given appropriate reading instruction at a young age, then special education services would not be necessary. He contended that research for SFA has shown that special education placement and referrals have been decreased after SFA was implemented.

Slavin, Madden, Dolan, and Wasik (1996) also evaluated SFA and control students in Baltimore. The results indicated that only two percent of the SFA school was reading below grade level as compared to nine percent of the control group.

A more current study was done using the data from the Baltimore school district. Borman and Hewes (2002) investigated the long-term effects and cost-effectiveness of SFA on the district. Using computerized data, they analyzed information from the Pupil

Information File and two standardized tests. The CAT was used as pretest reading data, and the CTBS was used as the eighth grade measure for math and reading achievement. The sample included the five original SFA schools and their matched control groups. In summary, the results indicated that the SFA eighth graders had higher reading ( $ES = +.29$ ) and math scores ( $ES = +.11$ ) than did their control group peers (Slavin & Madden, 2003). Also, they spent fewer years in special education ( $ES = +.18$ ) and fewer retentions ( $ES = +.39$ ) than the control group. Borman and Hewes (2002) also concluded that, like other elementary programs, the advantages of the program decrease over time. Although the SFA students did better than the control group, their scores compared to national norms were low.

Another district that has been involved in longitudinal studies is in Memphis, Tennessee. In 1988, Ross and Smith (1996) were hired to evaluate reading programs to increase the reading scores in Memphis since they had historically low achievement scores. After their investigation, they advised the district to institute SFA. The program started in 1990-1991 in one school, and then three years later in three other schools. By 1998-1999 school year, there were 40 schools in Memphis using SFA.

This district was comprised of 70% low-income students, 80% of whom were African American. Control schools were matched on these qualities as well as results from the CAT. Posttesting data were collected on individually administered reading tests as well as the Tennessee Comprehensive Assessment Program (TCAP), a part of the Comprehensive Test of Basic Skills (CTBS). Each school was followed individually over time. Results were given in effect sizes for individual schools and as grade levels.

In summary, results (Ross & Smith, 1996) from the individually administered reading tests indicated the largest effect size was seen in kindergarten with a  $+0.65$ . First grade had an effect size of  $+0.21$  and second grade had an effect size of  $+0.20$ . Although these results indicate a small effect size (AFT, 1998), Ross and Smith (1996) argued that the results are educationally significant. They did concede, however, that these results weren't conclusive of the programs effectiveness. They also contended that the small sample size contributed to the results. Only students who started and remained in the program were included; since the district had high mobility, this limited the number of participants.

Some researchers (Pogrow, 2000b; Jones et al., 1997) assert that SFA's policy of only including SFA students that were in the study from the beginning is problematic. Since SFA schools are usually in high poverty districts with high mobility rates, the program should be able to adjust to this fact.

According to Jones et al. (1997), Ross and Smith got some of their positive results by excluding two control groups. The control groups were dropped because the researchers believed that the schools were using some of the SFA components in their reading program. Since SFA used best practices for reading, Jones et al. argued that they were dropped for having effective reading strategies. Excluding them made the results more favorable to SFA. Other researchers have accused SFA of purposely including and excluding control and SFA groups to get the desired effect sizes (Venesky, 1998; Pogrow, 2000b).

Interestingly enough, there were no differences in the TCAP scores for the SFA and control schools. Ross and Smith (1996) interpreted this as the control school

teaching to the test, something that the SFA schools could not do since they had rigorous curriculum constraints. The SFA program has now been adjusted to correspond with standardized testing and/or state curriculums.

A somewhat different kind of study was done during the years 1995-1997 (Sanders, Wright, Ross, & Wang, 2000). Slavin, Madden, Dolan, & Wasik (1996) considered this one of the most important studies of the program. It incorporated the Tennessee Value-Added Assessment System (TVAAS) that was a program (Sanders et al., 2000) developed and analyzed by William Sanders at the University of Tennessee and was designed to help reduce biases that are believed to be inherent in standardized test scores. The TVAAS “gives an expected gain, based primarily on poverty levels, and compares it to actual scores on the TCAP” (Slavin, Madden, Dolan, & Wasik, 1996, p.29). The TVAAS scores for eight SFA schools were compared to control schools and to all Memphis schools. Grades included were from third to fifth.

Pretest scores showed that the SFA schools had lower TVAAS scores in reading, language, science, and social studies than the other two groups. Within two years, they scored significantly better than the other groups. This study indicates that the SFA program that focuses on reading, writing, and language arts can affect other subjects' scores. It also was the first time state assessments were used to evaluate SFA (Slavin & Madden, 2003; Slavin & Madden, 1996; Ross & Smith, 1996).

In the third year of this study, the SFA schools were compared to other comprehensive programs (Co-nect, Accelerated Schools, Audrey Cohen College, ATLAS, and Expeditionary Learning). The study looked at pre-reform TVAAS scores and compared them to post-reform scores. SFA showed not only the greatest gains but

also the highest scores. (Slavin & Madden, 2003; Slavin & Madden, 1996; Ross & Smith, 1996). The SFA schools had an average increase of 20 points over the control school, which is considered both statistically significant and educationally significant (Sanders et al., 2000).

In another study involving the Memphis school district, researchers (Cooper, Slavin, & Madden, 1998) compared 1998 reading and language achievement gains in 12 SFA schools. Expert trainers assessed the quality of implementation at these schools; they determined whether they were high or low implementation schools. These schools were compared to control schools (n=30) and others (n=28).

The high implementation schools scored higher than the comparison schools by 12 to 21 points but the results were not significant. The low implementation schools did the same or did only a little better than the control or other schools. When pre and post scores were analyzed, the high implementation schools had larger effect scores for both reading and language (+.97 and +1.22) when compared to the control schools. In language the high implementation schools had an effect size of +1.12 over the other schools.

Pogrow (2000a) later confronted SFA researchers, Slavin in particular, about these results. He focused on the TVAAS and its own ranking on schools based on the gains they had made. He stated that although SFA is used in 36% of the schools, 52% of those were placed in the lowest 77 of the state.

Nunnery, Slavin, Madden, Ross, Smith, Hunter, and Stubbs (1996), using quasi-experimental design, evaluated the effect that the degree of implementation had on reading achievement scores. The setting of this study was Houston, Texas. SFA was

instituted in 1994 in 50 schools. The schools were able to choose the degree of implementation; that is, the full SFA program, just the reading component, or the reading component with the tutoring. There were 23 schools from the district that comprised the control schools. The research was conducted in 1995 when the schools were in their second year of SFA. The pretest was the Language Assessment Scales. As a posttest, the high, medium, and low implementation schools' individually administered tests (Woodcock and Durrell) were compared to the control groups. Results were  $ES = +.47$ ,  $+.31$ , and  $-.13$ , from highest to lowest implementation. Overall, the higher the implementation, the greater were the achievement scores.

Additionally, schools that were predominately African American had higher scores than those that were mostly Hispanic. There was really no difference between the high and medium implementation for the Hispanic population. Second year results were better than first year results. The researchers concluded from the data that a full-time facilitator was an essential component in a high poverty and predominately African American school to ensure a raise in achievement level (Nunnery et al., 1996)

Pogrow (2002) claims that senior administrators in the Houston district commissioned their own evaluation that was never released. The three-year study included ten fully implemented SFA schools that were matched to 10 comparison schools. The results indicated that the SFA students had lower passing rates on the TAAS every year of the study, 1996-1998.

Slavin and Madden (2003b) cite preliminary results of a longitudinal study being conducted in Houston called Project GRAD. This program implements SFA in its entirety, as well as adds a math program, Move-It Math, and a school climate program,



Consistency Management/Cooperative Discipline. Research compared Project GRAD schools (n=8) with SFA schools (n=46) in Houston as well as with other Texas schools. TVAAS data indicated that the Project GRAD schools gained significantly more than SFA schools which did better than the other Texas schools.

Miami Dade County Public School incorporated SFA in 1992-1993 in two schools. Thirty-seven schools were using SFA by 1998. Urdegar (2000) conducted a research project to examine the impact of the program on the school district. During the 1998-1999 school year, he selected 18 schools based on similarity of ethnicity, English proficiency, and free/reduced lunch. Nine of the schools were SFA schools. Some of the SFA schools were also using technology components; three were using the Computer Curriculum Corporation (CCC), and four were using the Jostens Learning Centers. Nine schools were used as comparison schools; three schools were using the SRA/Reading Mastery, and six schools were using the district's Comprehensive Reading Plan. Urdegar's research included the level of implementation. Through surveys given to the administrators, teachers, and tutors and site visits by the SFA staff, he concluded that the school employees had the perception that they were fully implementing the program, but the SFA staff concluded that the school was only fully implementing three of the seven components.

Using the Stanford Achievement Test's (SAT) Reading Comprehension Scale as a comparison, Urdegar's (2000) research found no statistically significant differences in any of the programs. Later, in a Wall Street Journal editorial (Urdegar, 1999), he referred to SFA as unethical since it was actually detrimental to disadvantaged students.

Slavin and Madden (2003b), however, have taken exception to these results. They contend that by Urdegar's own admission the program was not implemented as it should have been. In addition, they state that Urdegar used flawed analysis by using covariate analysis as a pretest. They claim that since the SFA program was already in existence, this program already had an influence, so the results would be skewed. In addition, Slavin and Madden (2000) contend that the poor results were because the program was not implemented properly, the school system had a change in the superintendent, and the staff did not support the program. Pogrow (2000a) counters this by stating that a loss of superintendent is a common occurrence in urban districts and this should not affect the implementation of the program.

Ross, Smith, Casey, Johnson, and Bond (1994) compiled data in four cities using SFA. They evaluated one school in Memphis, Tennessee; two schools in Fort Wayne, Indiana; four schools in Montgomery, Alabama; and two schools in Caldwell, Idaho. They followed the Memphis school from 1990-1993 and the other schools from 1991-1993). They used the same type of quasi-experimental method with matched control group that other SFA studies had used. The researchers used MANOVA as their means of comparison. Reading achievement was measured by using the Woodcock and Durrell tests. Results are given in effect sizes.

Consistent with the other studies that Memphis was involved in, some of the results favored SFA. In the third year of the study, analysis showed an overall effect size of .+51 for the second graders who had had SFA since kindergarten. Other effect sizes were non-significant, but the SFA students did do better than the control group. The lowest achieving students had the largest benefits (Ross et al., 1994).

Results for the Montgomery schools were relatively weak the first year. Due to these results being inconsistent with other SFA studies, Ross and his colleagues made a follow-up visit the next year. Based on this observation, they contended that the control school was using a lot of the same strategies as the SFA school (i.e., tutoring, reduced class size, reading strategies). The results had improved in the second year of the study. In fact, the first graders did very well ( $ES=+1.32$ ) as did the lowest 25% performing students ( $ES=+2.86$ ) (Slavin & Madden, 2003a).

The Fort Wayne study (Ross et al., 1994) showed positive effect sizes on the reading comprehension of students as seen by the Indiana State Test for Educational Progress (ISTEP), the state achievement test. The effect size for students overall was  $+0.49$  and for the lowest 25% was  $+1.13$ . Second grade showed an effect size of  $+0.64$  and third grade was  $+0.13$ .

Caldwell, Idaho was the first rural SFA school to be studied (Ross et al., 1994). The students in this study, although having one of the highest increases in SFA schools, did not do any better than the control school, even though the program was being fully implemented. A team of researchers (Slavin, Madden, Dolan, Wasik, Ross, & Smith, 1994) contends that this was due to the control school being of high quality.

A statewide study of SFA was done in Texas using state data from the Internet (Hurley, Slavin, & Madden, 2001; Slavin & Madden, 2003a). This study was an attempt to address the problem of selection bias and to gather information on the effects on different ethnicities. All SFA schools from 1994-1998 were included ( $n=111$ ). These schools were compared to all the other schools in Texas using the available data from the Texas Assessment of Academic Skills (TAAS).

Special education students were not included in this testing until 1999, so comparisons of test scores could not be made before this date. Overall, the SFA schools were mostly Title 1 schools with high poverty levels and high minority population compared to the rest of the state. Results were given for grades 3, 4, and 5. On average, SFA schools made higher gains than the other Texas schools (ES+.59 or a gain of 5.85 percentage points). The authors commented, however, that a ceiling effect on the TAAS may have affected the scores.

Due to low sample sizes, studies in the past could not be done on the effects of SFA as they pertain to ethnic groups (Hurley et al., 2001). This study made an attempt to look at this variable. According to the data collected, both African American and Hispanic students in SFA made strides in closing the gap with white students. SFA African American students went from 63.3% passing in 1995 to 86.2% passing in 1998, the results for African Americans in the state went from 64.2% to 78.9%. Hispanic students gained more than the control group for three out of the four cohorts.

There is some question of the validity of SFA having an impact on closing the achievement gap (Pogrow, 2002). Pogrow argued that there were other factors that contributed to this increase in minority scores on the TAAS. Changes in promotion standards, lower class sizes, and test-prep efforts are a few statewide changes that may have contributed. He also contended that the study purposely compared SFA minorities to the state, instead of control schools, to achieve the desired results.

A similar study was done in California using the SAT-9 (Slavin & Madden, 2003b). All SFA schools that initially implemented the program in 1998 or 1999 were included (n=92) and compared to the state as a whole. Data were analyzed for these

schools from pre-implementation to spring, 2001. Combining both SFA cohorts, the differences were statistically significant ( $ES=+.25$ ,  $p<.02$ ).

Another study that was concerned with standardized test scores and ethnicity was conducted in two schools in a mid-western city (Ross, Smith, & Casey, 1997). It examined individually administered reading tests and mandated state tests on three grades over three years. A quasi-experimental design was used with two matched control groups. SFA was implemented in 1991-1992 due to low achievement of at-risk students. This setting was unique in that most SFA schools had a predominately minority population, and these schools had an almost equal mixture of white and African American students. In addition to the individually administered tests, such as Woodcock and Durrell, data were compiled on the state test, ISTEP.

In summary, the SFA students showed more gains on both the individual and state tests, with the exception of the third grade that was consistent with the control group. SFA students had a higher achievement level on the ISTEP. Previous studies (Slavin et al., 1994) had shown stronger results for the individually administered tests than the standardized tests.

Additionally, the results (Ross, Smith, & Casey, 1997) for ethnic group differences as they pertain to SFA were inconclusive. However, they do suggest that for first grade, the SFA minority students outperformed the control group; for second grade, there was no difference in the control and SFA group; and for third grade, the SFA group did slightly better than the control group. The authors theorize that this is because the low performing minority students were the ones that most benefited from the SFA components, such as the Family Support Team, tutoring, and reading strategies. Also, other

studies (Slavin et al., 1994) had shown that there is a benefit for the lowest performing 25%.

Another interesting outcome of this study (Ross, Smith, & Casey, 1997) was that the data indicated that there were decreasing scores over time. This also has been a conclusion of other research (Pogrow, 20001a, 2002; Ross et al., 1997, Ruffini et al., 1992). The authors argue that varying implementation quality may be a factor for this result (Ross et al., 1997).

Individual districts, that are not part of a formal study, send their own evaluations to the SFAF (Slavin et al., 1994). These testimonials have become part of the summary of research. For instance, Charleston, West Virginia reported that they had substantial gains on standardized test scores, increased their attendance, and had no retentions at all. Medesto, California increased their scores on the Comprehensive Tests of Basic Skills. A school in Wichita Falls, Texas claimed they increased their third grade passing scores from 48% to 70% in reading and from 8% to 53% in writing while the rest of the district stagnated.

In an attempt to overhaul the failing educational system and improve achievement scores in Charleston, South Carolina, SFA was instituted in one school in the district. This program became part of a three-year independent research study done by Jones et al. (1996). SFA was mandatory for this particular school due to its need for educational improvement; it was not voted on and did not receive the required 80% teacher approval rate that was usually necessary. It also did not include the family support component that would constitute full implementation. A comparison school was also chosen based on its

need of educational improvement. However, the demographics and educational levels were similar.

SFA was originally used in the 1989-1990 school year. The independent researchers used the similar format for evaluation as did previous SFA researchers. They used pretest information to match students, included only those students who completed the entire program, and used the individually administered tests for posttest data. SFA researchers were involved in this part of the study. Jones et al. (1995), however, also included standardized test score data since this was the data the school district wanted compiled. Math data was kept as well, since the district was concerned that the focus on reading would have an adverse effect on the math program. The results, calculated by using MANOVAs, were given for individual years and combined for the three years. In general, the data indicated that SFA had a statistically significant impact on the kindergarten students. The other grades did not produce the same results. In addition, the math scores generally went down for all grade levels.

Implementation of the SFA program at this school was problematic. As indicated earlier, this program was mandatory for the school district. According to Jones et al. (1996), it met with some resistance and interpersonal difficulties. However, no quantitative data were kept by either these researchers or the SFA researchers on the level of implementation. Possibly confounding the study was the fact that Hurricane Hugo hit in the beginning of the study, closing school briefly and causing some damage to the building.

Slavin and Madden (2003a) have commented about this particular study. They contend that there was poor implementation, little staff support, and a hurricane that

affected the results of the study. They claim that despite this, kindergarten and first grade measures showed that SFA had an impact. They conceded that the second and third grade measures did not indicate that SFA had an impact.

Jones et al. (1997) argued in their report that they did not think the hurricane affected the study since there was not any difference in the positive outcomes than the other years of the study. Pogrow (2000a) added to this observation that the control school was also affected by the hurricane so it would not taint the data.

Slavin and Madden (2001b) have compiled results from many of the SFA studies done in what they refer to as a multi-site replicated experiment. As Slavin (Slavin, Madden, Dolan, & Wasik, 1996) explains, “small-scale experiments located in different sites over extended periods are combined into one large-scale experiment” (p.198). They took all the data from all the schools for each grade in a given year. They have analysis for grades 1 to 5, with follow-ups for sixth and seventh grade. Included in the study were 23 SFA schools in 9 districts in 8 states. They reviewed the reading outcomes based on individually administered reading tests and changes in effect sizes over the years of implementation. The research concluded that significant effects were not seen for every grade level or in every district, but in general, positive effects were seen across the board.

Data indicated that the lowest 25% of each grade for the SFA schools had statistically significant differences (ES= +1.03 for first grade to ES= +1.68 for fourth grade). Students in SFA overall outperformed the control group. Interestingly, the mean effect sizes increased every year of implementation. In the follow-up portion of the study, the research indicated that, although the SFA students did better than the control students, neither group showed much growth in the middle grades.



A random study (SFAF, 2002) involving SFA schools is currently being done by the SFAF. It is a three-year, federally funded study being done in conjunction with the National Opinion Research Center. The research is being done on 41 Title 1 elementary schools. These schools have a poverty rate of 78% and are primarily in urban or poor rural areas. Half of the schools will randomly be assigned implementation of SFA in K-2, and the other half will implement SFA in grades 3-5. The opposite grades in each school will act as the control groups. The research is focused on reading and language skills, special education placements, grade retentions, disciplinary actions, and retentions.

The RAND Corporation (Berends, Kirby, Naftel, & McKelvy, 2001) has been accumulating data on the New American Schools since 1991. Although it funded 11 designs originally (Viadero, 2001), in 1995 it narrowed the number to seven when it actually began the programs on a large scale. Achievement outcome data were not studied until 1995; SFA data have been included since this time (Berends et al., 2001). In general, they have found that of the districts they have tracked (n= 163), 46% have had an increase in reading scores over the district scores. There was great inconsistency in the scores for each program. SFA (n=21) had the most consistency with 48% of the schools fairing better than their peers in the other district schools.

An open debate in the media has been taking place between SFA proponents and opponents. Each has blamed the other for biases that affect both opinion and research. Main opponents are Richard Venskey (1998), Stanley Pogrow (2000a, 2000b, 2002) and Rebecca Greenberg and Hebert Walberg (1998). All have questioned the methodology and samples of the SFA studies.

In general, the opponents have asserted that there are problems with the selection of the sample/control group, the data they choose to collect and report, and the fact that there are few third-party evaluations. Although SFA cited 24 independent evaluations (Slavin 1998a, Slavin & Madden, 2001b), Pogrow (2000a) stated that there really were only two independent evaluations. He stated that the Memphis study (Sanders et al., 2000) which Slavin and his colleagues emphasized as an independent review was not one in actuality, because Sanders and Wright were working in conjunction with the research group from SFA at their southern headquarters at the University of Memphis.

In addition, Pogrow (2000a) has questioned why Slavin (Slavin & Fashola, 1998) frequently has been given federal funds and government permission to evaluate his own and other programs, resulting in a biased assessment. Greenberg and Walberg (1998) have conferred that this is unsuitable. Pogrow (2000a, 2000b) also pointed to the one guide that seemed to be objective, *A Consumers' Guide to School Reform* (Herman, 1999). Although he does not name the specific researcher, he stated that there was one that had an affiliation with SFA.

Slavin has retaliated by pointing out potential biases of his critics (Slavin & Madden, 2000; Slavin, letter). He argued that Stanley Pogrow is the founder of a program named HOTS (Higher Order Thinking Skills) that was developed to enhance learning for disadvantaged Title 1 students after the third grade, and this program does not fit into the schoolwide reform model. He contended that because of this, Pogrow has tried to discredit the entire school reform movement. In fact, Pogrow (1994) has also found fault with other schoolwide programs. Pogrow has stated in his discussions that he believes

that these programs that include the entire school environment are hurting the Title 1 students.

Slavin (1998b) has made an attempt to discredit Greenberg and Walberg by indicating that they are trying to undermine Title 1 programs in an attempt to promote vouchers. Greenberg and Walberg (1998) responded by stating that their article was actually about research bias in program evaluations, and they were just presenting examples of this occurring in present day studies. They reiterated the fact that they thought there were serious flaws in the SFA research.

As a result of the Abbott decision, many schools in New Jersey had to choose a whole school reform. One district in a small urban area chose SFA in Spring 2000 after a vote from the district's teachers. Several programs gave presentations during in-service days during the school year. Those programs represented were Modern Red Schoolhouse, High Schools That Work, Co-nect, and Success for All. The vote for SFA was 80%.

SFA was instituted in the 2000-2001 school year in the elementary schools, Pre-K to third grade, and the middle school, fourth to sixth grade. The following year, 2001-2002, SFA was implemented in the seventh and eighth grades at the junior-senior high school. Funding for SFA came from Title 1 monies and CSRD monies (P. Claghorn, personal communication, October 30, 2003). Title 1 did not pay for SFA directly; it paid for teachers and support staff. The district also received over \$150,000 for professional development from Title 1. From CSRD, each school received \$50,000 a year for three years. The CSRD money was primarily used to pay for materials, books, and partial payment of the facilitators' salaries. Each school received an additional \$5,000 a year for

three years for administrative costs from CSRD. At the high school, the CSRD money was shared with the math program in 2003-2004, the last year of that school's three years. The cost of the SFA curriculum for this district was not available to this researcher despite several requests to the administration.

The implementation of the reading component (B. Norcross, personal communication, September 23, 2003) of SFA changed the scheduling of the school day. All classes at the elementary and middle school had reading during the same time period. The fact that the seventh and eighth grades were in the same building as the senior high (who were not mandated to have a WSR at this time) presented a scheduling difficulty. As a result, the junior high school students received SFA in 90-minute blocks, but the times varied for various groups. Students were grouped homogeneously based on SFA instituted testing. Special education students, except those who were being instructed in the self-contained classrooms, were included in the grouping. As in the other SFA programs, students were reassessed every eight weeks and regrouped, if necessary.

Teachers for SFA were trained for three days in the beginning of the school year of 2000-2001. Throughout the school year, there were several opportunities for additional training. All teachers were trained for the elementary and middle schools, but eight teachers at the high school (three of them special education teachers) were assigned to teach SFA. Tutors were also hired to assist in the reading component of SFA. In addition, extra aides were hired due to the amount of preparation of materials needed initially.

The Family Support Team was instituted at the schools during their first year of SFA (B. Wagner, personal communication, November 18, 2003). Each team included

the vice principal, the SFA facilitator, the school social worker, the school nurse, a guidance counselor, and the district's social services coordinator. Even though SFA was instituted in 2001-2002 at the junior high school, the training for this team was not fully implemented until 2002-2003. The team initially took over the Pupil Assistance Committee's (PAC) responsibility of handling student referrals and then added student and parent consultation and intervention to their role.

Due to the NCLB legislation, Abbott decisions, and the comprehensive school reform movement, New Jersey has been focused on raising test scores and achievement levels for disadvantaged students. The use of standardized testing to measure educational achievement for all students has roots in the Public School Education Act that was passed by the NJ legislature in 1975 (NJDOE, 2000b). An amendment was made in 1976 that set a minimum requirement for students' skills in communication and computation. This also made the passing of a standardized test mandatory for graduation; this type of test was given the first time in 1985-1986 to ninth graders and has gone through various changes throughout the years.

In 1996 the Core Curriculum Standards were passed which defined the skills the state deemed necessary for graduation (NJDOE, 2000b). The tests were adjusted to be in alignment with these standards. March of 1999 was the first time the Grade Eight Proficiency Assessment (GEPA) was given; this test replaced the Early Warning Test (EWT) that had been given since 1991. The EWT, and now the GEPA, is a tool schools use to determine which students are in danger of failing the high school test that determines graduation eligibility. The eleventh grade test was first given as the HSPT9

(High School Proficiency Test), then the HSPT11, and in 2000 the HSPA (High School Proficiency Assessment).

Recently, (Pearson Education, 2003) Pearson Educational Measurement received a two-year, \$11 million contract from the New Jersey Department of Education to continue to administer the GEPA. They are responsible for test assembly, psychometrics, publishing, distribution, scoring, and score reporting. The GEPA tests over 150,000 students every year. This company also provides assessment services for twenty other states.

The use of mandatory testing is controversial even though it is a requirement under NCLB. Due to NCLB, all states now have testing (USDOE, n.d.a). Several studies have expressed concern about this trend. The Civil Rights Project at Harvard University (Harvard, 2000) has been involved in two studies, both having to do with education reform. One study considered the influence that standardized testing had on instruction and curriculum. The study indicated that class time was spent on test-taking skills and practice tests at the expense of subject matter. Hence, high poverty schools teach the test and not the subject matter and strategies that relate to real-life settings. The authors concluded that testing took the focus and funding from schools and hurt the disadvantaged students. The second study evaluated state testing on ethnic and racial groups in Texas, New York, and Minnesota. The data indicated that the minority groups had lower scores than their non-minority peers in every case. The Civil Rights Project concluded that no single test should be used as a graduation requirement.

There have been lawsuits against this idea of judging a student by one criterion, the state exam. In 1997, the Mexican American Legal Defense Fund (Wildasky, 1999)

sued the Texas Education Agency for discrimination due to the fact that the majority of students who are denied graduation due to test failure are Mexican American and African American students. In 2000, the courts determined that the state exam was not discriminatory. (Lawsuit, 2004).

Currently, there are lawsuits against NCLB and state education agencies because of this idea of mandatory testing. There is a class-action suit against the Alaska Board of Education (Lewin, 2004) that contends that the state exam needed for graduation discriminates against disabled students. This suit was precipitated by the fact that more than 50% of the disabled students would be ineligible for graduation and federally guaranteed accommodations were not available. Another lawsuit is pending in Massachusetts (Lawsuit, 2003) about their state test, the Massachusetts Comprehensive Assessment System (MCAS). This lawsuit claims that the test is discriminatory against minority students. It also claims that this type of state test goes against the Education Reform Act of 1993 in that it is judging students on one assessment and that subjects other than math and English are being shortchanged.

Due to inadequate funding and the intrusive nature of NCLB (Toppo, 2004; Dobbs, 2004; Prah, 2004), many states are refusing to implement some or all of the mandates. Utah and Vermont have decided to only implement NCLB where there is enough federal funding. Arizona and Minnesota introduced legislation that would allow them not to fully implement NCLB (Dobbs, 2004). A total of 20 states (Toppo, 2004) are in the process of trying to amend or not comply with NCLB. Many of the issues involved impact testing and assessment issues.

## Chapter 3

### Methodology

This research was designed to evaluate the effectiveness of SFA on the raising of test scores in a particular low-income district. (The GEPA is the test used in the state of New Jersey to measure achievement levels of eighth graders in the core content standards. It is also the test used to determine if students in the Abbott districts are achieving as well as their non-Abbott peers).

Research design used for this study was quasi-experimental; the study contained a control group and an experimental group. The control group consisted of past eighth grade students who took the GEPA in 2001 who did not have SFA. This group was compared to the experimental group who were eighth grade students who took the GEPA in 2003 but had SFA for three years. The independent variable was SFA, while the GEPA was the dependent variable.

### Sample and Participants

The P-12 grade school district is located in a small urban community in southern New Jersey. This blue-collar town has a population of 11,484 (ELC, n.d.a). Enrollment in the schools is approximately 2,100. The district is homogeneous with 98% of the students being white; the average population for the Abbott districts is 13% white. In New Jersey schools as a whole, the average population of white students is 61%. Close to half of all students in this district are eligible for free/reduced lunch. Compared to the state average of 28% being eligible for free/reduced lunch, this is high; however, the average of all Abbott districts is 73%. Statewide, the average of students being classified as needing special education services is 15.7%; the average for Abbott districts is 16.3%.



In comparison, this district has a somewhat higher percentage of classified students at 22.1%. The average property value of homes is \$69,832. The state average property value is \$147,475; the average for Abbott districts is \$80,906.

A total of 550 students were involved in this study. The sample was one of convenience; no random assignments were possible. The experimental group (G1) consisted of the eighth grade class in the 2002-2003 school year. This class had SFA for three years, starting in sixth grade. In March 2003, this class took the GEPA. A total of 129 students took the test; 110 were general education students, and 19 (14%) were special education students. The control group (G2) was the group of students who took the GEPA in 2000-2001, before SFA was instituted at the junior high level. For this group, 133 students took the test; 95 were general education students, and 38 (28.5%) were special education students.

#### Instrument

Since New Jersey mandated WSR in the Abbott districts as a means to strengthen test scores in those schools, this research examined the GEPA scores as the dependent measure. The GEPA scores of eighth graders who had SFA and a different aggregate of students from the same district that took the test before the mandate of WSR were compared. (The GEPA is comparable from years 1999-2003 [H. Zhao, personal communication, October 13, 2003]. Reliability for the GEPA was .88 for the language arts/literacy section using Cronback's alpha). GEPA scores for the years 2001 and 2003 were analyzed to determine if SFA did have an impact on the language arts/literacy scores of students in the same district over time. This is a measure that the state and

federal government use to determine if a school district is improving. It is also a measure that SFA has used to determine its effectiveness.

The GEPA (NJDOE News, 2002) consists of three sections: science, mathematics, and literacy/language arts. The test is administered over a four-day period in March. The science portion has 60 multiple choice questions and 4 open-ended questions; the test takes 1 hour and 57 minutes. Day two of the testing is the mathematics portion. This test takes 2 hours and 27 minutes; there are 40 multiple choice and 8 open-ended questions.

The language arts/literacy portion is a two-day process. The first portion has ten multiple choice questions, two open-ended questions, and one writing task and is two hours and ten minutes long. The second portion of this section has the same format of questions but is administered in two hours and two minutes. There are three categories of scores; they are advanced proficient, proficient, and partially proficient. The scale score range is 100-300 (Zhao, personal communication, November 16, 2003).

#### Procedures

The data were examined by analyzing proficiency percents and mean totals for both years involved in the study. G1 and G2 were compared by looking at both the language arts/literacy of the GEPA to determine if there was a difference in scores after the institution of SFA. In addition, the groups were broken down into subcategories of special education and regular education to determine the impact on these subgroups. Also, an analysis was conducted on the number of students who scored advanced proficient, proficient, and partial proficient to evaluate whether there was a change in this type of data. Since SFA purported that their program increases the scores of the lowest

25% of the student body, data were analyzed for the lowest 25% of the GEPA scores for each year.

In addition to analyzing the proficiency rates, data were analyzed using total scale score means for each group to determine if the total mean improved or not. The difference between the mean score in 2001 and then in 2003 was determined.

However, the just proficient mean scores were different for each testing year and could not be compared. To determine if there was any difference in how close to the state mean each group was, the state just proficient means of each subsection of the language arts/literacy portion of the GEPA was compared to the total population and the subcategories of special education and regular education students in the district. This distance from the mean would determine if the groups were getting closer or further away from the state mean.

To determine if there was any change in the lowest 25% of the groups for each year, the individual scores for these students were averaged and the range was established. The two years were compared to measure how much closer they were to the proficient score of 200. Also, these scores were ranged from highest to lowest to see if there was a change from 2001 to 2003 in the range of scores.

The data were obtained through the district's high school guidance department. The results of the test are collected annually by the state of New Jersey and annually kept on file by the school district. Data were presented as individual and group data. In addition, the data on file included scores on subgroups of special education and regular education students. Permission to analyze the data was given by the principal of the school (Don, personal communication, June 18, 2003), the director of special education

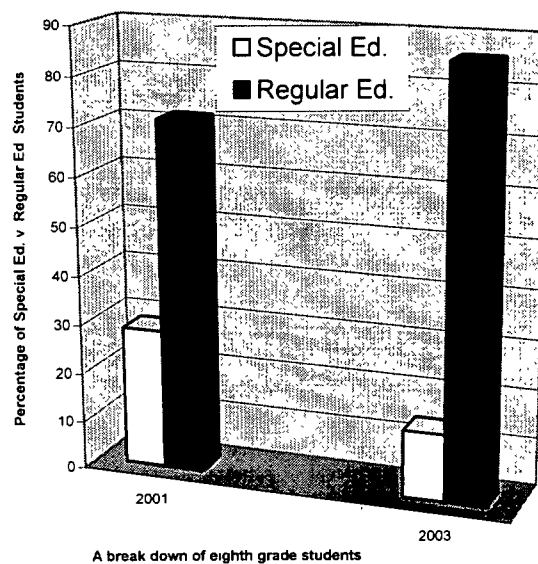
(Longer, personal communication, June 23), and the director of guidance (Koza, personal communication, September 4, 2003).

## Chapter 4

### Results

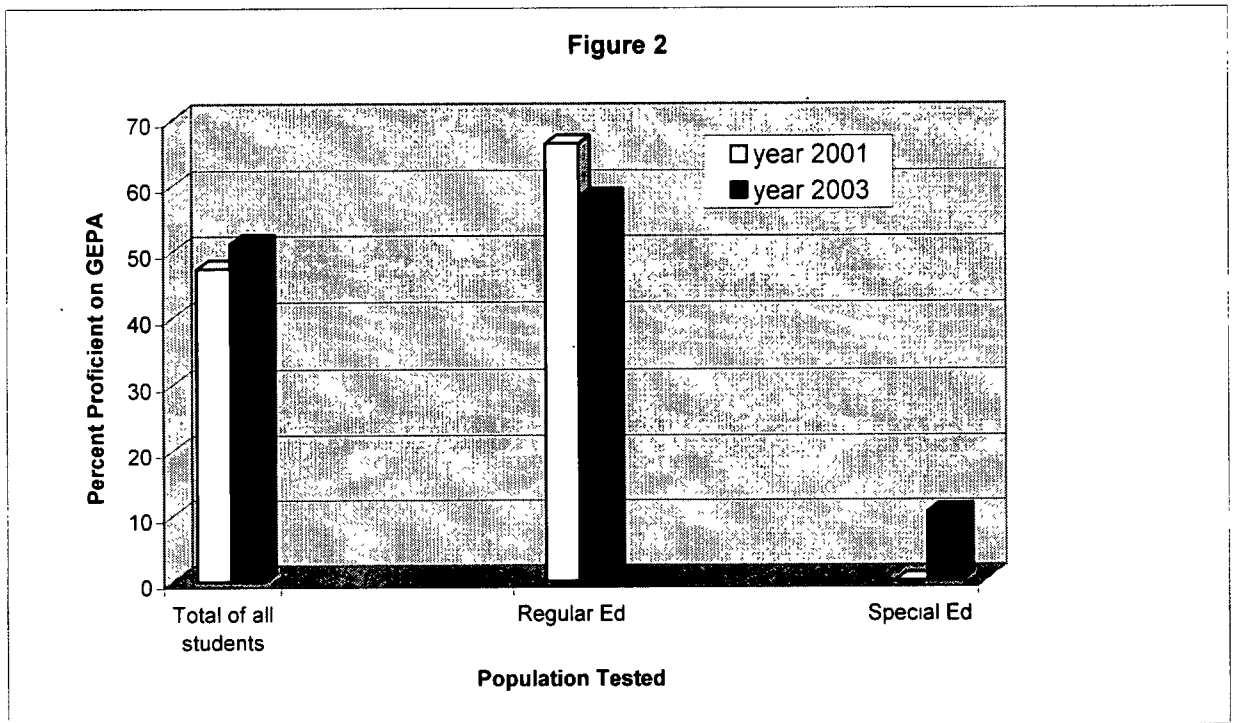
The percentage of regular education and special education students did change from 2001 to 2003. In 2001, 71.5% of the students were regular education students with 28.5% having a classification that placed them in special education. In 2003, only 14% of the students received special education services, leaving 86% of the student body identified as regular education. The special education population decreased 14.5% during this time period. See Figure 1.

Figure 1



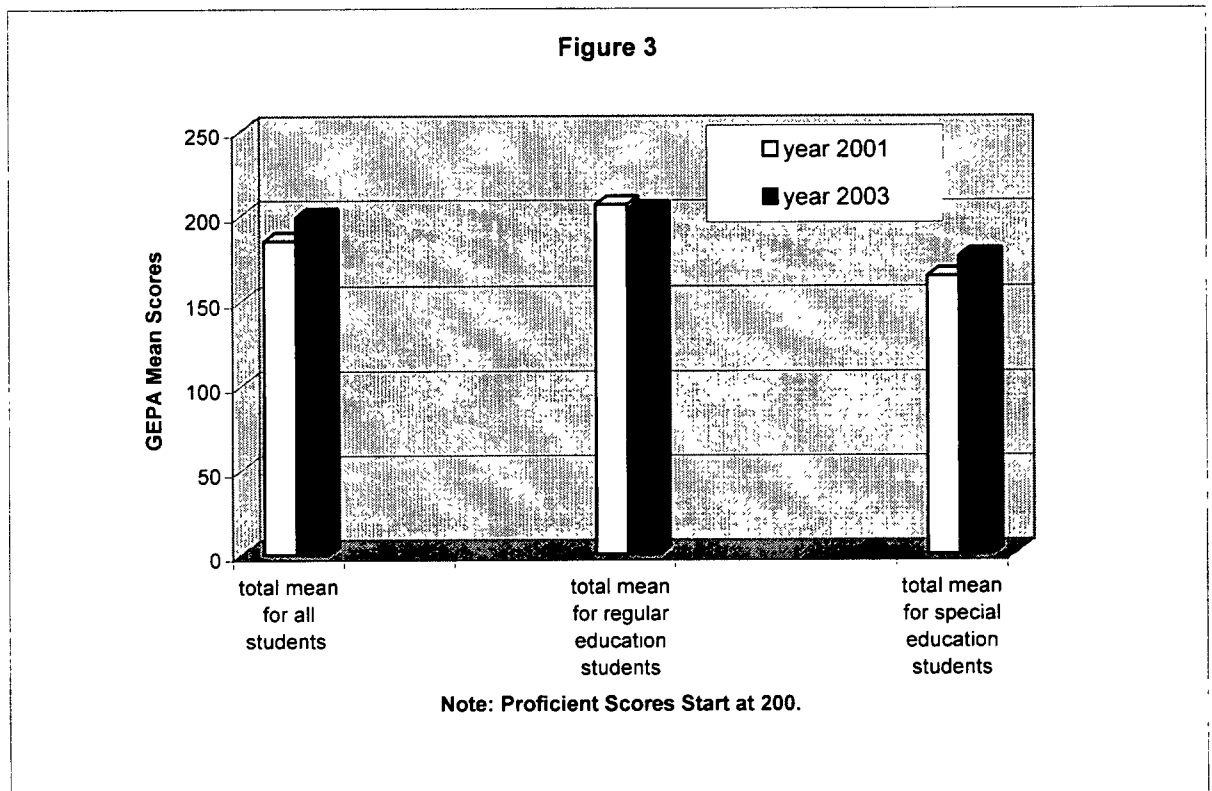
The results for the language arts/literacy portion of the GEPA for 2001 and 2003 showed some differences. Data indicated that of the total student population 47% were proficient in 2001. After SFA was implemented, 51.2% were proficient, a percentage

increase of 3.8. When the data were broken down into subgroups, it can be seen that 66.3% of the regular education students were proficient in 2001, and in 2003 there were 58% proficient, a decrease of 8.1%. On the other hand, special education proficiency went from 0% being proficient in 2001 to 10.5% becoming proficient in 2003. No students from either the 2001 or 2003 testing years achieved advanced proficient in the language arts/literacy section. The results are shown in Figure 2.



As can be seen in Figure 3, similar results were seen when the total scale score means for both years were compared. In 2001, the total scale score mean for all students was 185.25 as compared to 199.1 in 2003. When general education student scores are examined, the mean score is 206.1 for 2001 compared to 203.1 for 2003. Special education students had a mean increase of 11.8 points, from 164.4 in 2001 to 176.2 in 2003.

The language arts/literacy portion of the test contains four different sections: writing, reading, interpreting text, and analyzing/critiquing text. Although the just proficient means are different for each year of the study and can't be compared, data were analyzed to determine how far from the mean regular education and special education student scores were. The mean scores of each group were compared to the just proficient mean of that year to determine how far away from the mean they were (positive or negative distance). This number was then compared to the same type of data for 2003. In this way, it could be determined if the groups got closer or further away from the mean.



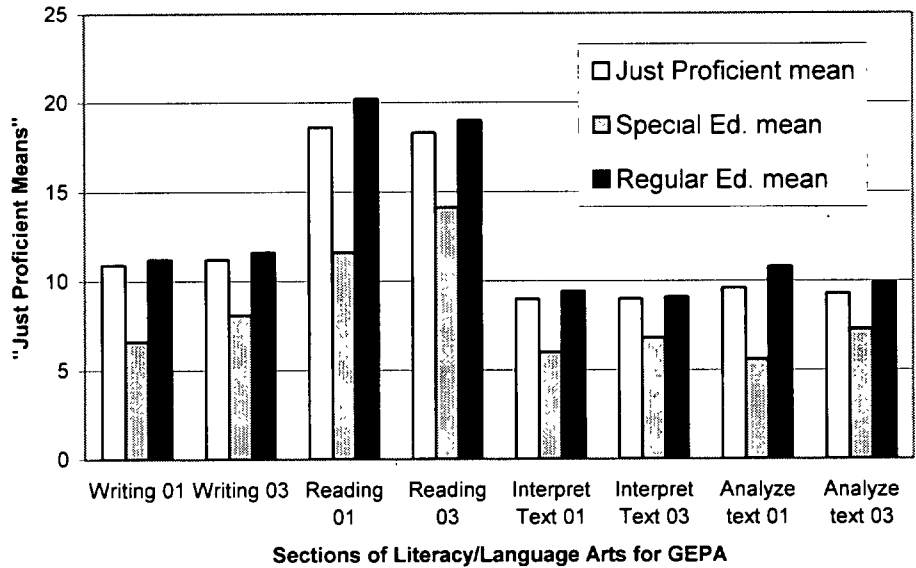
In general, the regular education scores remained above the mean while special education scores approached closer to the mean in the 2003 data. See Figure 4. Although the general education scores remained above the mean, overall scores decreased.

The writing section was the only section in which this population showed an increase; the scores went from .3 above the mean to .4 points above the mean. This indicated an increase of only .1 points. The reading score for this population went from 1.6 points above the mean in 2001 to 1.3 points above the mean in 2003, a decrease of .3 points. Interpreting text also showed a decrease in .3 points from 2001 and 2003; these scores went from .4 above the mean to .1 above the mean. The analyzing text scores for 2001 were 2.8 points above the mean; however, this score fell to .4 above the mean, a decrease of 2.4 points. Overall, this population lost a total of 2.9 points.

Special education data indicated that this group's scores grew closer to the mean in 2003 in all sections. As evidenced by closer mean scores, reading was the section of the test in which this group showed the most improvement; they went from 7 points below the mean to 4.2 points below the mean, a difference of 2.8 points. The section of the test that evaluated analyzing text indicated a difference of 2 points, from 4 points below the mean in 2001 to 2 points below in 2003. Data for the writing section indicated a 1.2 difference, 4.3 points in 2001 as compared to 3.1 in 2003. Interpreting text had the smallest degree of change for this group; they were 3 points below the mean in 2001 and 2.2 points below the mean in 2003, a .8 difference. This population had an overall increase of 6.8 points.



Figure 4



When analyzing the lowest 25% of the scores on the GEPA for language arts/literacy, there is an improvement in scores from 2001 (n=33) to 2003 (n=32). The range of scores for the lowest 25% is 131 to 173 for 2001 and 132 to 185 for 2003. Data were analyzed to determine the mean for this group for each year. The mean of the lowest 25% of the scores was 156.7 and 170 for 2001 and 2003, respectively. This was an increase of 13.3 points after SFA was instituted.

## Chapter 5

### Discussion

There are several claims that the creator of SFA has made that seem to be validated by the findings of this study. The achievement scores of the total population of the eighth grade in this Abbott district, as evidenced by the percentage of students who were proficient on the GEPA, did increase after SFA was incorporated in the junior high. An increase in mean scores for this group also showed improvement. This increase in achievement scores has been reported many times by SFA and outside sources (Slavin, Madden, Dolan, & Wasik, 1996; Hurley et al., 2000; Ross, Smith, & Casey, 1997; Slavin & Madden, 2003).

However, achievement did not increase for all groups. In fact, general education students actually did not do as well on the state test after they were exposed to SFA as they had done previously. This finding is supported in the literature on SFA. Independent studies in Houston (Pogrow, 2002) indicated that the state test results after SFA was implemented actually went down each year. Other studies concluded that the SFA schools did not do as well or did the same as the control schools on state testing (Ross et al., 1994; Ross & Smith, 1996; Pogrow 2000a).

On the other hand, more classified students scored proficient on the GEPA after the implementation of SFA. This group also had an increase in overall mean scores for the test, as well as scores closer to the mean on subtests. In addition, the population of special education students decreased. As Slavin (1995) has indicated, the intense reading instruction incorporated in SFA should make it so placement and referrals to special

education would not be necessary. It is also a policy of SFA not to classify students (Slavin & Madden, 2001b).

This decrease in special education placement may have had an impact on the GEPA scores for both regular and special education students. Because the population of classified students was almost cut in half, those students who would have been tested under the auspices of special education were tested as regular education students. This redistribution of students may have been responsible for lowering the test scores of the regular education group.

It is possible that the decreased number of special education population is a result of non-identification of potential special education students and not an actual decrease in special education needs. As Ruffini et al. (1991) indicated, this policy of not classifying students could lead to lower regular education test scores. If potential special education students were not absorbed in the regular education data, then it seems that the intense reading program was not beneficial to everyone, particularly regular education students.

Furthermore, the increase in special education proficiency was a result of 2 out of 19 classified students passing the GEPA in 2003. No special education student had passed the 2001 GEPA. Small sample sizes in this study had a great impact on the statistics.

When comparing the lowest 25% of the scores, there is an increase in scores after SFA. The mean score for this group increased 13 points. The lowest test score in 2001 was 173 and this low score was raised to 185 in 2003. This again indicated that this group is getting closer to reaching the proficiency cut off of 200. This is consistent with

SFA reports of an increase for this group (Slavin, 1994; Slavin, Madden, Dolan, & Wasik, 1996).

Some factors that must be considered when discussing this study are those that opponents and proponents have argued about in the literature. This study did not account for how long students were exposed to the SFA program. Ideally, the students would have been in the district for the three years SFA had been implemented. This district (NJDOE, n.d.c) had a mobility rate of 21.1% in 2001 as compared to a rate of 8.2% in 2003 (the state average was 11.9 and 10.9, respectively). This indicates the student population was more stable during the implementation of the whole school reform. SFA reports do not include students who were not in the program the entire time. But as Pogrow (2000b) and Jones et al. (1997) contended, high poverty districts have mobile students and any program that purports to assist them should be able to accommodate this factor.

The idea of implementation should also be addressed. This district was not a fully implemented SFA program. Not all aspects of the program were incorporated initially; they evolved over the three years. It was not considered a fully implemented school by SFA (B. Norcross, personal communication, September 23, 2003). As other studies have indicated (Slavin & Madden, 2000; Nunnery et al., 1996; Ross et al., 1997), this may have impacted this district's GEPA scores.

This district, as part of the Abbott legislation, was mandated to have whole school reform. Although there was a majority vote of 80% agreement by the staff, it was a concept forced upon them. This, too, may have been a factor on how well the program

was implemented. Studies (Slavin & Madden, 2000; Jones et al., 1997) have shown that reluctant staff could affect the outcomes of the program.

Although some studies show an increase in certain scores of SFA schools over control schools (Slavin & Madden, 2001), some researchers (Borman & Hewes, 2002; Venesky, 1998) have argued that the achievement scores are not close to national or state norms. In this particular study, the results seem to suggest that there still exists an achievement gap. For example, the special education percent that was proficient in 2003 was 10.5; this can be compared to the state average of 27.7% (NJDOE, n.d.c). SFA proponents ( Ross & Smith, 1996; Hurley et al., 2001) would argue that this was due to the fact that there were already very low tests scores, so any improvement was significant. However, when the percent proficient is compared with other districts that have similar socio-economic levels and historically low test scores, this Abbott district still falls short; the average adjusted by economics for these other districts is 15.8% (NJDOE, n.d.c).

Along this line of thought, it is important to consider the regular education percent proficiency in the same way. For 2003, 58.3% of the regular education population was proficient; the state average for this population was 73.8%. Districts that have similar poverty levels as the one examined had 64% of the regular education proficient. It is also interesting to analyze the language arts/literacy GEPA test scores for the regular education population over time for the district studied. The percent proficient in 2001 was 66.3%, in 2002 was 60.3%, and in 2003 the proficiency percentage was 58.2. The students in 2001 did not have SFA at all, the students tested in 2002 had one year of SFA, and the students in 2003 had 3 years of SFA. This shows a consistent downward trend in

state test scores the longer students were exposed to SFA. This coincides with the results of Pogrow (2000a, 2002) and Ross et al. (1997).

A consideration needs to be made as to whether any of these changes in the test scores are significant for educational purposes. This study shows mixed results for the success of SFA on the tests scores of this district. It is not clear, though, that these results are enough to make a judgment on their impact. Is an increase of 3.8% on the total population enough to justify the cost? Was it even significant that the special education population had an increase of 11 mean points when the majority of these students were still not proficient? If the bottom 25% gets closer to the mean at the expense of the top 25% improving, is this a worthwhile program? Is the achievement gap being closed by keeping the top down? These are all questions that must be pondered.

Moreover, there are many limitations to this study that must be considered. The changes in test scores cannot be determined to be a direct result of SFA. As Pogrow (2002) indicated, other factors may have been going on in the school that contributed to the changes. Increased emphasis and stress on these standardized scores by the state and national government may have been a factor. In addition, any school policy changes may have influenced the students during this time period.

Another problem with this study is the evaluation technique that was utilized. The students who were tested in 2001 were an entirely different cohort than the ones who took the test in 2003. They took two different tests in two different years. Circumstances and ability levels for the groups may have been markedly different. Comparing two different GEPA tests can also be problematic. Although the tests are evaluated for validity and reliability, the questions are different and certain scores cannot be compared

for different years (Zhao, personal communication, November 16, 2003). However, this is how the state of New Jersey does evaluate this data. They compare one year's scores to another to determine improvement. This is also how NCLB is evaluating annual yearly progress for school districts. If SFA is intended to improve state test scores, than this type of evaluation should be sufficient.

The sample size of the groups involved is also small. The special education population in 2003 was only 19 students. As indicated earlier, two students passing the GEPA created a 10.5% increase in this category. This is somewhat misleading. This is true for all the categories of students; a difference of one or two test scores could be perceived as an important change statistically.

Due to the limitations of the study and the sample sizes, there should not be any generalization about SFA and its impact on test scores. However, for this district, certain conclusions can be made based on the results of the testing. Further assumptions may have been made if it was possible to have a pretest available for both the control and experimental testing group. This extra information could have given insight into whether one group initially had different ability levels than the other. The information could have also assisted in evaluating whether there was a marked improvement in the same group from before SFA and after SFA. This would have eliminated some of the speculation involved in the study.

This study focused on students who did not receive SFA until they were in sixth grade. SFA by its own account (Slavin & Madden, 2000) prefers to start the program at the early elementary level. The initial goal of the program was to have all students reading on grade level by third grade. Since this group of students studied was not

brought all the way through the program, this may have had an impact on the results.

However, research has been reported on grade levels up to middle school with students who were not involved in SFA during their early elementary schooling.

Since most of the studies that are conducted on SFA included a control group, it would have been easier to refute or accept SFA claims if a control group was instituted in this study. As it was noted earlier, however, this district is unique in its racial and economic composition. In the final analysis, it may have been better to evaluate the program based on its claims to raise standardized test scores. This is, in fact, the intention of the Abbott legislation and the concept of whole school reform.

Further hypotheses that could have been addressed could have focused on the science and math scores of the GEPA. Did the emphasis on reading instruction in the SFA program detract or add to the other sections of the test? Claims have been made that SFA has increased test scores in all areas (Slavin, Madden, Dolan, & Wasik, 1996). Two other areas of education that SFA contends their program enhances are attendance and discipline referrals. If these factors were considered in this study, then it would have furnished another dimension as to the change in test scores. These two factors alone could have an impact on how well students perform in a testing situation. Being in school more days, may have contributed to their test-taking ability, despite the curriculum changes involved.

Further information and study needs to be done on all WSR models, not just SFA. The goal of making all students proficient is a noble one. The question still remains as to how to make this happen. The expensive WSR models that are being executed still remain controversial in their methods and effects. Independent research by



the districts that are using the methods may be the only way to truly determine what is effective and what is not. Even then, as was seen in this district's research on WSR and standardized test scores, the impact may still be inconclusive.

## References

- American Federation of Teachers (AFT). (1998). *Building on the best, learning from what works: Six promising schoolwide reform programs*. Washington, D.C.
- Berends, M., Kirby, S., Naftel, S., & Mcklevey, R. (2001). *Implementation and performance in New American Schools: Three years into scale-up*. Rand Corporation.
- Borman, G. & Hewes, G. (2002). The long-term effects and cost-effectiveness of Success for all. *Educational Evaluation and Policy Analysis*, 4, 243-266.
- Borman, G., Hewes, G., Overman, L., & Brown, S. (2002). *Comprehensive school reform and student achievement: A meta-analysis*. Baltimore, MD: Johns Hopkins University.
- Clune, W. (1994a). The cost and management of program adequacy: An emerging issue in educational policy and finance. *Educational Policy*, 8(4), 365-375.
- Cooper, R., Slavin, R., & Madden, N. (1998). Success for all: Improving the quality of implementation of whole-school change through the use of a national reform network. *Education and Urban Society* 30(3), 385-408.
- Crosbie, J. (2000, August 10). *Number of schools in New Jersey's Abbott Districts adopting whole school reform*. New Jersey Department of Education. Retrieved July 18, 2003, from <http://www.state.nj.us/njded/news/2000/0810wsr.htm>
- D'Agostino, Borman, Hedges, & Wong (1998). Longitudinal achievement and chapter 1 coordination in high-poverty schools: A multilevel analysis of the prospects data. *Journal of Students Placed at Risk* 3(4), 401-420.

Dobbs, M. (2004, February 19). More states are fighting 'No Child Left Behind' law.

*Washington Post*. Retrieved on April 12, 2004 from

<http://www.washingtonpost.com>

Doherty, K. (2000). *Early implementation of the Comprehensive School Reform*

*Demonstration (CSR D) program*. Jessup, MD: U.S. Department of Education,  
Editorial Publications Center.

Education Law Center (n.d.a) *Abbott district profiles*. Retrieved September 19, 2003

from <http://edlawcenter.org/ELCPublic/AbbottvBurke/AbbottProfile.htm>

Education Law Center (n.d.b). *About Abbott v. Burke*. Retrieved July 19, 2003, from

<http://www.edlawcenter.org/ELCPublic/AbbottvBurke/AboutAbbott.htm>

Education Law Center (n.d.c). *History of Abbott*. Retrieved September 19, 2003, from

<http://www.edlawcenter.org/ELCPublic/AbbottvBurke/AbbottHistory.htm>

Fashola, O Slavin, R. (1998). Schoolwide reform models. *Phi Delta Kappan* 7(5), 370-380.

Greenberg, R. & Walberg, H. (April 8, 1998). [Letter to editor]. *Education Week*.

Retrieved November 9, 2003, from <http://www.edweek.org>

Greenberg, R. & Walberg, H. (1999). The Dioceses factor. *Phi Delta Kappan*, 127-128.

Goertz, M. (1994). Program equity and adequacy: Issues from the field. *Educational Policy* 8 (4), 608-615.

Harvard University Gazette (2000, January 20). Studies: High stakes tests are

counterproductive to economically disadvantaged students. Retrieved October 8, 2003, from Harvard Gazette Archives. <http://>

[www.news.harvard.edu/gazette/2000/01.20/tests.html](http://www.news.harvard.edu/gazette/2000/01.20/tests.html)

- Herman, R. (1999). *An educator's guide to schoolwide reform*. Arlington, VA: Educational Research Service.
- Hurley, E., Chamberlain, A., Slavin, R., & Madden, N. (2001). Effects of Success for All on TAAS reading: A Texas statewide evaluation. *Phi Delta Kappan*, 82(10), 750-756.
- Jones, E., Gottfredson, G., & Gottfredson, D. (1997). Success for some: An evaluation of a Success for All program. *Evaluation Review*, 21(6), 643-670.
- Kantor, H. (1997). Equal opportunity and the federal role in education. In *Funding for Justice: Money, Equity, and the Future of Public Education*, Milwaukee, WI: Rethinking Schools, Ltd., p.69-76.
- Karp, S. (1997). Equity suits clog the courts: Legal battles for state remedies face limitations. In *Funding for Justice: Money, Equity, and the Future of Public Education*, Milwaukee, WI: Rethinking Schools, Ltd., p.5-9.
- King, J. (1994). Meeting the educational needs of at-risk students: A cost analysis of three models. *Educational Evaluation and Policy Analysis*, 16, 1-19.
- Lawsuit challenges MCAS. (2004, January 8). GazetteNet. Retrieved March 29, 2004 from <http://www.gazettenet.com/schoold/01082003/3377.htm>
- Lewin, T. (2004, March 17). Disabled Alaska students sue over exam. *The New York Times Archives*. Retrieved March 29, 2004 from <http://query.nytimes.com/gst/abstract.html>
- Lowe, R. (1997). Race, power, and funding: an historical perspective. In S. Karp, R. Lowe, B. Miner, B. Peterson (Eds.). In *Funding for Justice: Money, Equity, and*

*the Future of Public Education*, Milwaukee, WI: Rethinking Schools, Ltd., p.14-16.

Madden, N., Slavin, R., Karweit, N., Dolan, L., & Wasik, B. (1993). Success for All: Longitudinal effects of a restructuring program for inner-city elementary schools. *American Educational Research Journal*, 30,123-148.

The National Institute on the Education of At-Risk Students [NIEARS]. (1998, April). Success for All in *Tools for Schools*. Retrieved September 20, 2003, from <http://www.ed.gov/pubs/ToolsforSchools/sfa.html>

New Jersey Department of Education [NJDOE]. (2000a, January). *Historical overview of school reform in the Abbott or Special Needs Districts*. Retrieved July 18, 2003, from <http://www.state.nj.us/njded/abbotts/eval/shu/chap2.shtml>

New Jersey Department of Education (2000b). *Examiner manual: Grade Eight Proficiency Assessment: 2000 sample test*.

New Jersey Department of Education (n.d.). *2002-2003 school report card*. Retrieved April 14, 2004 from <http://education.state.nj.us/rc/roo3/index.htm>

New Jersey Department of Education News (2002, March 8). *GEPA testing begins next week*. Retrieved December 17, 2003, from <http://www.state.nj.us/njded/news/2002/0308gepa.htm>

Northwest Regional Educational Library (NWREL). (n.d.) Success for All/Roots & Wings. *The Catalog of School Reform Models*. Retrieved November 2, 2003, from <http://www.nwrel/scpd/catalog>

- Nunnery, J., Ross, S., Hunter, P., & Stubbs, J. (1997, March). *Effects of full and partial implementations of Success for All in English and Spanish*. Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL.
- Orthfield, G. (1994). Asking the right question. *Educational Policy* 8(4), 404-413.
- Pearson Education (2003, September 16). *Pearson wins two-year extension for New Jersey test*. Press Release. Retrieved November 9, 2003, from [http://www.pearsoned.com/pr\\_2003/091603c.htm](http://www.pearsoned.com/pr_2003/091603c.htm)
- Pogrow, S. (1994). A skeptical perspective on the adequacy conception. *Educational Policy*(8)4, 414-424.
- Pogrow, S. (2000a). Success for All does not produce success for students. *Phi Delta Kappan*, 67-79.
- Pogrow, S. (2000b). The unsubstantiated 'Success' of Success for All: Implications for policy, practice, and the soul of our profession. *Phi Delta Kappan*, 596-600.
- Pogrow, S. (2002). Success for all is a failure. *Phi Delta Kappan*, 83(6), p.463-469.
- Prah, P. (2003, July15). No child law could spawn state lawsuits. *Government Technology*. Retrieved March 29, 2004 from <http://www.govtech.net/news/news>.
- Ross, S. & Smith, L. (1996). Success for All in Memphis: Raising reading performance in high-poverty schools. In R.Slavin., N. Madden, L. Dolan, & B. Wasik, B. (Eds.). (1996). *Every child, every school: Success for All*. Newbury Park, CA: Corwin. (pp.49-78).
- Ross, S., Smith, L., Casey, J., (1997). Preventing reading school failure: Impacts of Success for All on standardized test outcomes, minority group performance, and

- school effectiveness. *Journal of Education for Students Placed at Risk*, 2(1), 29-53.
- Ross, S., Smith, L., Casey, J., Johnson, B., & Bond, C. (1994, April). *Using Success for All to restructure elementary schools: A tale of four cities*. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.
- Rothmiller, R. (1994). Equality or adequacy of school funding. *Educational Policy*, 8(4), 616-625.
- Ruffini, S., Feldman, B., Edirisooriya, G, Howe, L., & Borders, D. ( 1992). *Assessment of Success for All*. Baltimore City Public Schools.
- Sanders, W., Wright, S., & Ross, S., & Wang, L. (2000). *Value-added achievement results for three cohorts of Roots and Wings Schools in Memphis: 1995-1999 Outcomes*. Memphis: University of Memphis Center for Research in Educational Policy.
- Slavin, R. (1994). Statewide finance reform: Ensuring educational adequacy for high-poverty schools. *Educational Policy*, 8(4), 425-434.
- Slavin, R. (1996). Neverstreaming: Preventing learning disabilities. *Educational Leadership*, 53 (5), 4-7.
- Slavin, R. (1998a). Research overwhelmingly supports Success for All. [Letter to editor]. Retrieved November 7, 2003, from <http://www.successforall.net/resource/research>

- Slavin, R. (April 29, 1998b). Slavin responds to essay's 'ad hominem' critique. *Education Week*. Retrieved November 9, 2003, from <http://www.edweek.com/ew/vol-17/33letter.htm>
- Slavin, R. & Fashola, O. (1998). *Show me the evidence: Proven and promising programs for America's schools*. Thousand Oaks, CA: Corwin.
- Slavin, R. & Madden, N. (1999). *Disseminating Success for All: Lessons for policy and practice*. Success for All Foundation. (Report No. 30). Retrieved November 7, 2003, from <http://www.successforall.net/resource/research/scalingup.htm>
- Slavin, R. & Madden, N. (2000). Research on achievement outcomes of Success for All: A summary and response to critics. *Phi Delta Kappan*, 82(1), 38-40, 56-66.
- Slavin, R. & Madden, N. (2001a, April). *Reducing the gap: Success for All and the achievement of African-American and Latino students*. Paper presented at the meeting of the American Educational Research Association, Seattle, WA.
- Slavin, R. & Madden, N. (Eds.). (2001b). *One million children: Success for All*. Thousand Oaks, CA: Corwin.
- Slavin, R. & Madden, N. (2003a). *Success for All/Roots & Wings: Summary of research on achievement outcomes*. John Hopkins University, Center for Research on the Education of Students Placed at Risk (CESPAR).
- Slavin, R. & Madden, N. (2003b). *Summary of research on achievement outcomes*. (Report No. 41). Center for Research on the Education of Students at Risk (CESPAR).
- Slavin, R., Madden, N., Dolan, L., & Wasik, B. (1994). Whenever and wherever we choose...the replication of Success for All. *Phi Delta Kappan*, 75, 639-647.



- Slavin, R. & Madden, N., Dolan, L., & Wasik, B. (Eds.). (1996). *Every child, every school: Success for All*. Newbury Park, CA: Corwin.
- Slavin, R., Madden, N., Karweit, N., Livermon, B., & Dolan, L. (1990). Success for all: First-year outcomes of a comprehensive plan for reforming urban education. *American Educational Research Journal* 27(2), 255-278.
- Springfield, S. (2000). A synthesis and critique of four recent reviews of whole-school reform in the United States. *School Effectiveness and School Improvement* 11(2), 259-269.
- Success for All Foundation (SFAF). (2002, September 12). *Landmark study to determine school reform model's impact*. Retrieved September 19, 2003, from <http://www.sucessforall.net>
- Toppo, G. (2004, February 11). States fight No Child Left Behind, calling it intrusive. *USA Today*. Retrieved April 12, 2004 from <http://www.usatoday.com/news/education>
- Traub, J. (1999). *Better by design? A consumer's guide to schoolwide reform*. Washington, D. C: Thomas Fordham Foundation.
- Urdegar, S. (1999, August 10). Success for all is unethical [Letter to the editor]. *Wall Street Journal*, p. A-1.
- Urdegar, S. (2000). *Evaluation of Success for All Program, 1997-1998*. Miami: Office of Evaluation and Research, Miami-Dade County Public Schools.
- U.S. Department of Education. (n.d.a). *Achieving equality through high standards and accountability*. Retrieved October 5, 2003, from [http://www.ed.gov/nclb/overview/intro/presidentplan/page\\_pg4.html](http://www.ed.gov/nclb/overview/intro/presidentplan/page_pg4.html)

- U.S. Department of Education. (n.d.b). *Preliminary overview of programs and changes included in the No Child Left Behind Act of 2001*. Retrieved October 5, 2003, from [http://www.ed.gov/nclb/overview/intr/prgsum/sum\\_pg2.html](http://www.ed.gov/nclb/overview/intr/prgsum/sum_pg2.html)
- Venesky, R. (1998). An alternative perspective on Success for All. In K. Wong (Ed.), *Advances in educational policy, vol. 4* (pp. 145-165). Greenwich, CT: JAI Press.
- Viadero, D. (2001, April 18). RAND finds mixed results for school reform models. *Education Week*. Retrieved November 7, 2003, from <http://www.edweek.org>
- Walker, E. & Gutmore, D. (2000). *The quest for equity and excellence in education: A study of whole school reform in New Jersey special needs districts*. New Jersey Department of Education: Abbott Implementation. Retrieved September 9, 2003 from <http://www.state.nj.njded/abbotts/res/shu/>
- Wang, M., Haertal, G., Walberg, H. (1997). *What do we know?* Philadelphia, PA: Temple University Laboratory for Student Success.
- Wildasky, B. (1999, September 27). *Achievement testing gets its day in court paired: A key Texas case goes to trial as minority groups seek to stop a school reform trend*. U.S. News and World Report. p.30-32.

Appendix A

IRB Letter

2004-60

RECEIVED FEB 04 2004

Appendix C

INSTITUTIONAL REVIEW BOARD  
DISPOSITION FORM

*Data Review*

Principal Investigator

Lisa Labbree

Address of Principal Investigator

88 Baynes Ave

City, State, and Zip Code

Gloucester City, NJ 08030

Telephone # Fax # e-mail address

(856) 456-5580 lisaghs@comcast.com

TITLE OF RESEARCH

The effects of Success for All (SFA) as a whole school reform on GEPA scores in an Abbott District

Co-Principal Investigator (if applicable)

Address of Co-Principal Investigator

City, State, and Zip Code

Telephone # Fax # e-mail address

ADMINISTRATIVE DISPOSITION - DO NOT WRITE BELOW THIS LINE

Your claim for exemption for the research study identified above has been reviewed. The action taken is indicated below:

APPROVED FOR EXEMPTION AS CLAIMED: CATEGORY # \_\_\_\_\_

Note: Anything that materially changes the exempt status of this study must be presented to the IRB for approval before the changes are implemented. Such modifications should be sent to the IRB Office at the address above.

APPROVED FOR EXEMPTION - BUT NOT AS CLAIMED. Your claim for exemption does not fit the criteria for exemption designated in your proposal. However, the study does meet the criteria for exemption under CATEGORY # \_\_\_\_\_.

A determination regarding the exempt status of this study cannot be made at this time. Additional information is required.

Your proposal does not meet the criteria for exemption, and a full review will be provided by the IRB.

EXPEDITED REVIEW:  Approved  Denied

FULL REVIEW:  Approved  Approved with modifications  Denied

DENIED: \_\_\_\_\_

See attached Committee Action Letter for additional comments.

Chair, IRB

Date

*J. Guak*

*2/20/04*

Co-Chair, IRB

Date

*[Signature]*

*2/18/04*

Appendix B  
Approval Letter

# Gloucester City Junior-Senior High School

1300 Market Street

Gloucester City, New Jersey 08030

(856) 456-7000 • Fax: (856) 456-2348

Jack L. Don - *Principal*

Dennis Perry - *Assistant Principal*

Robert MacCausland - *Assistant Principal*

Peter Koza - *Director of Pupil Personnel*

January 6, 2004

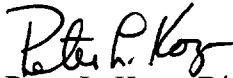
To whom it may concern:

This letter is to verify that Lisa Labbree received the results of the GEPA for school years 2001, 2002, and 2003, from the guidance department at Gloucester City High School.

She has permission to use these results and the supporting data in her master's thesis.

If you have any questions regarding this matter you may contact me at (856) 456-7000, extension 1545.

Thank you,



Peter L. Koza, Director of Guidance

