The Effect of a State Department of Education Teacher Mentor Initiative on Science Achievement

by

Stephen L. Pruitt

A dissertation submitted to the Graduate Faculty of
Auburn University
in partial fulfillment of the
requirements for the Degree of
Doctor of Philosophy

Auburn, Alabama
August 9, 2010

Keywords: Science Education, Coaching, Mentoring, Inquiry, Economically Disadvantaged

Approved by

Carolyn S. Wallace, Chair, Associate Professor of Secondary Science Education
Charles J. Eick, Ph.D., Associate Professor of Curriculum and Teaching
David M. Shannon, Ph.D., Professor of Educational Foundations, Leadership, and Technology
Farooq A. Khan, Ph.D., Professor of Chemistry

Abstract

This study analyzed a state department of education’s ability to have actual influence over the improvement of science achievement and proficiency by having direct relationships with science teachers in Georgia’s lowest performing schools.

The study employed a mixed ANOVA analysis of the mean scale scores and proficiency rates of the science portion of the Georgia High School Graduation Test (GHSGT) for the years 2004 through 2007 to determine if the intervention by the Science Mentor Program (SMP) had significant effect on the science achievement and proficiency within the cohort of schools, as compared to a set of schools receiving no intervention, on various subgroups within the schools, and on various levels of intervention within the SMP. All data used in this study are available to the public through the Georgia Department of Education (GaDOE). SMP schools were selected based on their level of intervention for three consecutive years. Non-SMP schools were selected based on demographic similarities in economically disadvantaged, white, African-American, and students with disabilities to ensure a match of pairings for analyses.

The results of this study showed significant improvement of scale scores and proficiency rates between 2004 and 2007. The study showed significant increases in all schools regardless of treatment. The study also showed significant differences in performance within the subgroups. Males, white, non-Economically Disadvantaged, and regular education students were all found to have significantly better performance in both achievement and proficiency rate. Economically Disadvantaged students were found to have a significant difference with regard to treatment groups. There was a significant difference between the mean scale score and proficiency rates of

Economically Disadvantaged students in schools receiving high-intervention and schools receiving no-intervention. Further analysis showed that the only significant difference was in 2004, the year prior to implementation. Results indicate while the high-intervention schools did perform lower over all four years, they were not significantly different during the time of treatment indicating high-intervention schools performed at levels equivalent to schools receiving no-intervention.

This study provided evidence of the success of a specific intervention by a state education agency to improve science education for the practicing teacher and its role in improving student science achievement. It will be used by policymakers to determine future activities and potential funding of other such programs. This also has a potential for national use as it is the only program of this nature operated by a department of education in the country.

Acknowledgements

The author wishes to express his appreciation to those who supported him through his goal of receiving his Doctorate of Philosophy. The most important person the author wishes to thank is his wife, Cecelia Pruitt. Without her constant support and push, this milestone would never have been achieved. There are not enough ways to express the gratitude or admiration due to such a remarkable lady. In addition, without the support and understanding of his children, Samuel and Abigail, he would not have had the strength to continue. While a doctorate has always been a goal of his, it pales in comparison to the opportunity to show his children the benefits of hard work and family support. The author wishes to thank family, who through care for the children, comforting smiles and encouragement, allowed him this opportunity; his parents, Howard and Jeannie Pruitt, and his aunt and uncle, Tom and Joyce Foster. Special acknowledge must go to James and Wilma Cannon. My only regret is that they could not be here to see me finish. I do believe they know how much they helped me.

Finally, the author wishes to thank the members of his committee, each of whom played key roles in his success. First, thanks to his major professor, Carolyn Wallace. Dr. Wallace was extremely helpful and supportive in helping the author navigate through his dissertation. Thanks, too, go to Dr. Charles Eick who helped and supported the author through residency and oral examinations and introduced him to a whole new academic world. Special appreciation is also given to David Shannon and Farooq Khan. Dr. Shannon graciously agreed to participate on the author's committee very late in the process and was instrumental in the interpretation of the analyses. Dr. Khan, who serves as a Professor of Chemistry, at the University of West Georgia, graciously agreed to travel to Auburn to support the author and represent the world of chemistry.

Table of Contents

List of Tables

List of Figures

Chapter I. Introduction

Improving science achievement is a concern for many states in the United States. This study will analyze a state department of education's ability to have actual influence over the improvement of science achievement by having direct relationships with science teachers in Georgia's lowest performing schools. The Georgia Department of Education (GaDOE), science educators and policymakers, have intervened to provide continuous effort to improve and facilitate quality science instruction. This study will determine the level of success of a specific intervention to improve and enhance science education for the practicing teacher and its role in improving student science achievement.

Scientific Literacy

October 4, 1957 was one of the most significant days in American education. It was the day people in the United States knew they had a problem. It was not the Russians as many thought, it was the quality and rigor of science education in the U.S. The day *Sputnik* was launched the U.S. felt, and rightfully so, that the nation had fallen behind. As with most paradigm shifts, however, changes have been slow in American education. While the U.S. was able to reach the moon first, it became apparent that its youth needed quality science and mathematics education. Since 1957, there has been much research into science education and how to challenge our students appropriately. Studies such as the Trends in International Mathematics and Science Study (TIMSS), Program for International Student

Assessment (PISA) and the National Assessment of Educational Progress (NAEP), the only federally funded assessment, have shown the U.S. is still lacking in its ability to educate students in science education. While the U.S. is competitive at an international level at the fourth grade, we fall behind at the eighth grade, and drastically behind by the end of twelfth grade (USED, 2007).

The National Science Teacher Association (NSTA) in partnership with the American Association for the Advancement of Science (AAAS) has set the goal that all Americans will be scientifically literate by the year 2061, the next time Halley's Comet comes into view of the earth. In 1996, the National Research Council (NRC) targeted that effort by publishing the *National Science Education Standards*. Federal legislation known as *No Child Left Behind* (NCLB) has somewhat accelerated that timeline. Due to high stakes testing, NCLB reinforces the natural tendency of schools and school systems to concentrate on reading and mathematics; this reinforcement can be seen through the amount of funding given to reading and mathematics instructions (Wheeler, 2004). But in a society run by technology and science, emphasis must be placed on science as well. If the citizens of the U.S. want to remain at the top of technological countries in the world, we must train ALL of our students in the content and thinking skills of science (Page, 2004).

Additionally, a focus on STEM (Science, Technology, Engineering and Mathematics) has become punctuated by its position of importance in President Barak Obama's education platform. A scientifically literate person as defined by National Science Education Standards (NRC, 1996) is one who "can ask, find, or determine answers to questions derived from curiosity about everyday experiences" (p.1). A scientifically literate person can read scientific articles and discuss them in a public conversation, can identify scientific issues at local and national levels and communicate a position on such matters. Being scientifically literate does not mean a person can quote many facts about science or memorize the periodic table. That is to say, students are expected to understand and apply knowledge rather than memorize facts (NRC, 1996). Science is a way of knowing, not a way of remembering.

There has been greater emphasis placed on science instruction in recent years, largely due to needs in society with regard to economics and national security, by organizations such as the National Science Teachers Association (NSTA), American Association for the Advancement of Science (AAAS), Trends in International Mathematics and Science Study (TIMSS), and the National Research Council. Increasingly, more private and public organizations, such as the Gates Foundation and the United States Department of Education, have stated how critical it is that the United States stays atop the world with regard to science and mathematics. The 2005 National Assessment of Educational Progress indicated that fourth graders have shown a significant increase in performance both in scale score and proficiency since 1996. Eighth graders' performance remained steady since the 1996 assessment, but twelfth graders declined (NCES, 2000).

These data could indicate science teachers are not holding students' interest nor giving high school students the proper tools to be considered scientifically literate. In losing interest in science, students may opt out of chemistry, physics, and other higher level sciences, and thus decrease their ability to become full and productive citizens in modern society. There have been many hypotheses regarding poor science test results in the U. S., including problems with the curriculum, assessment, teacher qualifications, pedagogical methods and student attitudes, as well as social factors, including a vast diversity of ethnicities, socio-economic levels, cultures, and factors involved in children's lives outside of school (Wilson, Taylor, Kowalski, & Carlson, 2010; Penfield, & Lee, 2010). A closer look at the number of science courses required and qualified teachers in each state are also important components to observe. Standardized tests have often been guilty of assessing knowledge at the skills and recall level; however, this type of assessment or curriculum does not guide the instruction of scientific concepts toward understanding and higher level thought.

The NAEP assessments are effective at assessing a student's ability to process a problem and arrive at a conclusion. Released questions from the 1996 administration of NAEP required eighth graders to determine salt concentration using a floating pencil and twelfth graders to separate an unknown mixture. Tasks such as these require students to apply scientific knowledge and practice to arrive at a quality conclusion. Students must have an understanding of density and concentration, for instance, if they wish to understand the task. They must understand quality practice to experimentally determine the unknown salt concentration. They must display good experimentation skills as well as mathematical reasoning and graphing. All of this must be completed in 30 minutes time (National Assessment of Educational Progress, 2009). NAEP describes its science test in the following manner,

> *"Each exercise in the science assessment measures one of the elements of knowing and doing science within one of the fields of science (for example, scientific investigation in the context of physical science). In addition, one-half of the students in each school received one of three hands-on tasks and related questions. These performance tasks require students to conduct actual experiments using materials provided to them, and to record their observations and conclusions in their test booklets by responding to both multiple-choice and constructed-response questions."*

There is also much to be learned from the 1999 and 2007 TIMSS. The 1999 assessment showed that U.S. eighth graders ranked eighteenth among the thirty-four countries who participated (TIMSS). In 2007, U.S. eighth graders ranked eleventh out of forty-nine countries. While the achievement scores did increase and the U.S. was above the international average, the U. S. has a long way to go to reach its goal of scientific literacy for all. I am an advocate that the students of the United States deserve instructional methods that will give them a better chance to succeed. Given all the research on outcomes, why is it that some of our students still do not achieve in science?

The goal of scientific literacy does not limit itself to only a select group of students. All students need and should have open access to scientific processes and content. In *Science For All Americans*, Rutherford and Ahlgren (1990) state that all students can learn the standards defined in the document. In *Project 2061*, they suggest that Students With Disabilities (SWD) should be included as part of science education reform. Of course, the basis of this reform was to focus on understanding concepts and relationships as opposed to memorizing facts and vocabulary (1993). However, the reform has still shown significant differences in performance of various groups of learners. Minorities, SWD, and Economically Disadvantaged (ED) tend to score lower on standardized tests than whites, regular education students, and non-ED. In 2005, the National Center for Education Statistics reported that SWD consistently scored approximately one standard deviation below regular education students on the 2000 fourth-, eighth-, and twelfth-grades on the National Assessment of Educational Progress (NAEP) in science (Mastropieri & Scruggs, 2006). NAEP scores showed a decrease in the gap between ED students and non-ED at both fourth and eighth grades between 1996 and 2005. In 1996, fourth-grade non-ED had a higher scale score (159) and percent proficient and above (37%) than ED students (129, 12%, respectively). In 2005, while non-ED still outperformed ED, ED saw a significant increase in both scale score (158) while proficiency rate stayed steady (12%). Like overall students, no significant change was found in eighth grade ED versus non-ED (Grigg, Lauko, & Brockway, 2006).

Embracing the idea that science is a way of knowing from a policy or curricular standpoint might be easily accepted by stakeholders such as administrators, teachers, parents and students. However, our teaching workforce must not only embrace inquiry, but also be able to implement it and diagnose the misconceptions students have regarding science. For some teachers, science has remained a list of words or facts which exacerbates the problem with preparing students for scientific literacy and a standardized science test that includes problem solving and inquiry at its core. In schools that have traditionally shown low performance in science, there has been little in the way of a support system for the teachers to enable them to engage in inquiry-based teaching. To address this problem, the Georgia Department of Education sought to render that service in 2005 through a program known as the Science Mentor Program (SMP). The program is designed to be an intervention in schools exhibiting low performance on the science portion of the Georgia High School Graduation Test (GHSGT).

Georgia Implications

Accountability, high stakes testing, student achievement, and *No Child Left Behind* (NCLB) are all buzz words that every teacher, student or administrator will probably hear at least once in a typical school day. In order to meet societal demands for accountability in education, the Georgia legislature approved participation in the NCLB program. NCLB requires states to administer high stakes assessments to determine student achievement in order for a state to continue to receive federal funding. Effective educators are always looking for better, more effective ways to teach their students. Educators want what is best for students, and most want to find better ways to educate their students. In turn, they must provide data that indicate the "Annual Yearly Progress" (AYP) of student achievement in science.

Georgia has revised its standards. The new standards are called the Georgia Performance

Standards or GPS. The GPS emphasize scientific practice in the context of scientific content knowledge. The old curriculum in Georgia was originally written in 1985 and revised once in 1996. The old curriculum contained four "process skills." The first dealt with inquiry from a data collection and analysis point of view, the second dealt with researching media for use with a history context, the third was a safety objective, and the fourth dealt with natural resources or industry uses. These were largely ignored by many science teachers (GaDOE, 2002). The new GPS have been written in terms of a dual expectation. Georgia has adopted the *Benchmarks for Scientific Literacy* scientific practices and refers to those as Characteristics of Science. The terminology used is meant to show the significance of both science as inquiry and scientific content, the two co-requisites. Within the introductory paragraph of each grade level and course is the following statement:

> Science consists of a way of thinking and investigating, as well a growing body of knowledge about the natural world. To become literate in science, therefore, students need to acquire an understanding of both the **Characteristics of Science** and its **Content**. The Georgia Performance Standards for Science require that instruction be organized so that they are treated together. Therefore, **A CONTENT STANDARD IS NOT MET UNLESS APPLICABLE CHARACTERISTICS OF SCIENCE ARE ALSO ADDRESSED AT THE SAME TIME.** For this reason they are presented as co-requisites. (GPS, 2004)

The language is very clear as to the meaning of the dual requirements. The document was placed on the web for ninety days for public review and comment. One of the larger concerns dealt with the funding for this way of teaching (GaDOE, 2004). While this is a valid concern, the charge to the Department of Education from the State Board of Education was to deliver "world class" standards (GaSBOE, 2004). Without the emphasis on inquiry and processes, it could not be a world class curriculum. This was a paradigm shift from the past; it may be possible that many teachers do not understand the need for the dual expectations or how to implement them.

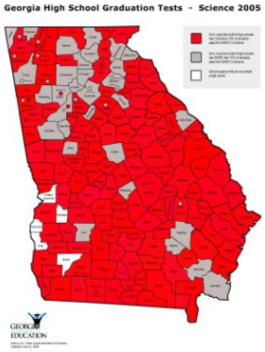
The GPS contain nine process standards. The first seven are the Habits of Mind; the last two are termed the Nature of Science. Habits of Mind deal with skills we want to see developed in all students. These standards deal with questioning/curiosity; estimation/computation; the use of tools and instruments, including technology; the application ideas of systems and models; communication/writing; reading across the curriculum; and scientific discernment. The Nature of Science deals with the character of scientific knowledge and important features of inquiry. Unfortunately, I believe there are many teachers and administrators who do not know the purpose of these standards and how to act upon them. Some comments during the public comment period showed concern over the placement of these within the document. The concern was that it sends the message that the two co-requisites should be treated separately.

The Characteristics of Science standards are also tested through the content on the Criterion Reference Criteria Test, given to students in third through eighth grade; the End-of-Course Test, given at the end of Biology and Physical Science; and the Georgia High School Graduation Test, given during the eleventh-grade year. Teachers who have worked on the curriculum revisions have made every attempt to remove fact-based items from these assessments and concentrate on processes, such as utilizing population data over a period of years to determine the relationship of various organisms in an ecosystem. The assessment is based on the new GPS, and so content and process will also be tested together. Items are being "double coded" to meet both the content standards and the Characteristics of Science standards. In the first year of students taking the transitional version of the Georgia High School Graduation Test, state scores increased by 5%. While this was a significant jump, there remained a crisis in Georgia as evidenced by the fact that only 75% of students passed the science portion of the GHSGT the first time.

In reality, improving science scores can be dealt with only through professional development and support. In 2003, only 70% of first time test takers were passing the GHSGT in science. By 2005, the passage rate had improved to 75% (GaDOE, 2005). Several factors that have led to improvements in science achievement in Georgia. The Georgia Department of Education requested 1.3 million dollars for use with the content specific training. Georgia is by no means the first to attempt the emphasis on process skills and inquiry. However, it is one of the first to require it as a co-requisite to the content taught (GaDOE, 2003). The universities and colleges in Georgia are committed to helping train future teachers in this method of teaching, but there is much to still be done. With the whole curriculum built on inquiry practices, more training and materials are crucial. There is more to do with less funding available. However, expecting effective and expedient implementation was not practical without more support at the building level. In 2005, the GaDOE implemented a program to support classroom teachers with the implementation of the new standards and inquiry.

Georgia administers two statewide assessments, the Georgia High School Graduation Test (GHSGT) and the End-of-Course Test (EOCT). The GHSGT fulfills the NCLB requirement while the EOCT fulfills Georgia House Bill 1187, also known as the A+ Reform Act. Both assessments contain scientific practices and content as equally assessed items. We have had a crisis situation with regard to our science achievement in Georgia. While the new standards and assessments require the use of scientific practices, many teachers were not fully prepared for this implementation. Many of Georgia’s lowest performing schools also had teachers who had not been engaged in inquiry-based learning or assessment. From 2003 until 2005, first-time test takers had a passage rate of 70%, 71%, and 71%, respectively. One hundred and forty out of 180 (78%) of school systems had less than half their high schools with passage rates greater than or equal to 70%. As shown in Figure 1 below, most of the state was in a crisis. The red school systems represent those with half or more of their high schools having less than 70% of their first-time test takers pass. Gray represents school systems that had greater than half, and the white are counties in Georgia that do not have a high school (GaDOE, 2005).

Figure 1 – 2005 GHSGT Science Performance Map



Georgia's predominantly red map visually displays the science achievement crisis in Georgia. In addition to stagnated science achievement scores, a new and more rigorous set of standards requiring the use of scientific practices (inquiry) was on the horizon. There was great concern that if the state's low performing schools were not able to improve achievement with what amounted to a discrete list of facts, how would they improve with a new set of standards on which a new test was under development that required the use of data and evidence? How could the new standards be implemented effectively? The typical professional learning activities were not effective. Teachers needed job-embedded professional learning, modeling, resources, and expertise not always readily available in rural areas. To truly improve classroom practice effecting student achievement and build capacity to facilitate quality practice, teachers need a network (Potter & Reynolds, 2002) and a program focused on teacher leadership (Elmore, 1996; Glickman, 2002; Gordon, 2004; Murphy, 2005).

> "Teacher leadership might be most valuable as a means to enhance the professional growth and development of teachers, … and their interactions with their colleagues in ways that enhance student learning and increase the capacity of the school to adapt and

> improve." (Conley, 1997)

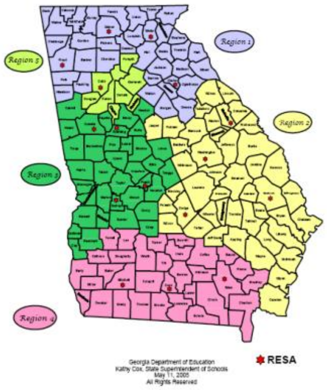
The answer came in the form of the Science Mentor Program (SMP).

Overview of the Science Mentor Program

The Science Mentor Program was conceived in the spring of 2005 as a result of the poor science achievement on the GHSGT by first-time test takers. The program was developed following the 2004 Georgia General Assembly allocating $2,000,000 with a specific charge to support classroom science teachers. The GaDOE developed a plan to place mentors in classrooms to improve instruction, enhance use of inquiry, and provide quality, job-embedded professional learning to science teachers. At the inception of the program, GaDOE made a conscious decision to employ currently practicing science teachers rather than administrators to ensure current best practice. They were also selected based on their understanding of and training on the new science GPS, inquiry, and pedagogical content knowledge. The Science Implementation Specialists (SIS) were trained in the philosophy of the new GPS, but had to show evidence in understanding inquiry methods. SIS were given the charge to mentor and coach struggling science teachers in the areas of content and inquiry pedagogy, build strong relationships with the science teachers in their area to provide ongoing support, act as a science liaison to the state, and perhaps most importantly, build capacity throughout the state by establishing teacher leaders within each school serviced by the program.

The SMP employs seventeen teachers located throughout the state. The state is divided into five regions. Each region has four members except for the fifth region, which employs one full-time teacher, but is supported by staff from the surrounding regions as shown in Figure 1.2 below.

Figure 2 – Science Mentor Regional Map

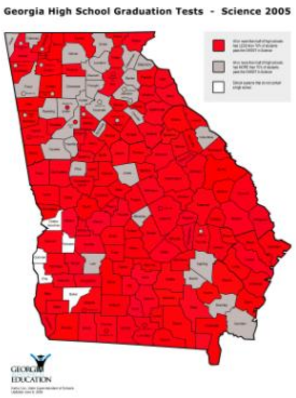


Early in the program, the decision was made to place staff in rural areas where teachers had the least support from their county office. In these areas, there are no specific individuals responsible for improving specific content areas. Whereas most of the metropolitan Atlanta area such as Gwinnett, Cobb, and Fulton counties have science supervisors who provide support and professional learning, these rural systems do not. They also have some of the poorest achievement results and represent some of the most economically challenged areas in Georgia.

Schools were selected using a formula that resulted in a need factor. The schools with the highest need factors were selected to receive the highest levels of support. The formula was based on five areas: 1) GHSGT Science Data; 2) EOCT Science Data; 3) Graduation Rate; 4) AYP Status; and 5) Number of students taking GHSGT. Need Factors are based on a weighted average of the five areas above with GHSGT being weighted most and the number of students taking GHSGT least.

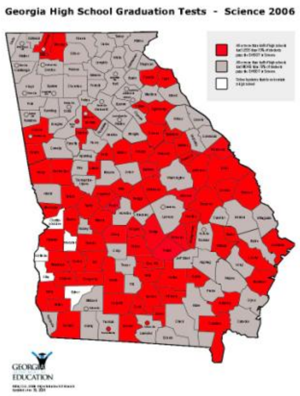
In the 2005-2006 school year, the first year of the program, SIS worked with 122 of Georgia's 375 high schools. In the first year, because many school districts only have high one school, the number of school systems on the science "red" list dropped twenty-five percent. The 2006 map (Figure 3) below has the same conditions as previously discussed.

Figure 1
2005 GHSGT Science Performance Map



2004-2005

Figure 3
2006 GHSGT Science Performance Map



2005-2006

Purpose and Research Questions

This study will analyze the effects of placing science mentors in schools that have traditionally exhibited low performance in the area of science on the GHSGT. Science Mentors delivered two types of interventions to these schools. High-level intervention involved a consistent presence within the school involving service at least one full day per week while medium-level service received service at least two times per month. Specifically, the researcher will examine two questions: 1) Is there a significant difference in science achievement and proficiency for schools supported by the SMP each year in performance, between SMP and comparable schools, and between SMP schools receiving medium- versus high-level intervention on the science portion of the GHSGT from 2005-2007 during the period of intervention? 2) Is there significant improvement in the science achievement and proficiency on the science portion of the GHSGT for subgroups [male, female; Economically Disadvantaged (ED), non-Economically Disadvantaged (non-ED); students with disabilities (SWD), non-SWD; White, Black, Hispanic, Asian] within schools receiving high-level intervention by the SMP and between SMP and comparable schools from 2005-2007?

The researcher will utilize a quantitative methods approach to the study in order to analyze the effects of the SMP. The GHSGT science achievement scores for first-time test takers in the initial cohort of schools selected for the SMP will be analyzed to see if student achievement in the SMP showed significant change from year to year. The study will focus on schools identified as medium level (SMP intervention at least two times each month) and as high level (SMP intervention at least weekly). The study compares year-to-year changes in achievement for SMP high-level intervention SMP schools, compared to a group of non-SMP schools and compared to medium-level intervention SMP schools. For the 49 original high-level intervention schools identified in the original cohort in 2006, a comparison set of schools with similar demographics will be selected to measure effectiveness of the treatments administered through the SMP. The schools will be selected based on similarities in population, economically disadvantaged (ED), and percent of students with disabilities (SWD). The study will focus on year-to-year change over a three-year period. The analysis will utilize a mixed ANOVA. This design should allow the researcher to focus on the improvement in science achievement scores for SMP schools as compared to non-SMP schools. The year-to-year improvement is the within effect for all of the research questions.

Significance of Study

The results of this study will be significant because it renders evidence policymakers need to improve science achievement. Improving science achievement has been an underdeveloped area on the part of most state departments of education. As science is beginning to take its place as an important component in a child’s education, teachers who have not been able to improve the achievement of their students need more support. Science scores in Georgia have traditionally had the lowest passing rate and the lowest first-time pass rate of the four content areas (English Language Arts, Mathematics, Social Studies and Science). It is the contention of the researcher that it is, therefore, necessary to provide ongoing, effective, and fiscally efficient, support to students and teachers in low-performing schools. This program is unique in its implementation when compared to other state science initiatives. It is important to evaluate the effectiveness of an intervention specifically designed to provide teachers with professional development and instructional support.

This study should also yield additional insight into teacher support, professional development, and the effect of changing teacher practice. Teachers whose students are struggling in science have been targeted for this intervention. Key factors used to try to change the results have been focused on site-based professional development with an emphasis on using inquiry in science instruction. As can be seen in the following chapter, literature shows on-site professional development is a key factor in improving or changing practice. This study will contribute to the growing research in the area of teacher professional development.

In addition, this study will contribute to research in the area of impacting achievement through high levels of intervention through mentoring and coaching strategies. The study will evaluate the effectiveness, based on standardized tests, of a program operated and implemented by individuals not employed as traditional teachers. That is to say, the ability of “outsiders,” such as coaches and mentors, to impact practicing teachers and their practices should give insight into the effectiveness of programs that employ mentoring practices.

Chapter II. Review of Literature

Science – A Way of Knowing

Science is a way of knowing. It is a way of understanding our natural world and the processes that occur within it. Since science is an endeavor to understand all that we can about the natural world, ongoing investigation and research must continue in order to shape and reshape our understanding. With each passing year, more discoveries sharpen our focus, thereby allowing us to clarify our perspective on how and why things work as they do. Scientists, in order to learn more about our world, must inquire about it. They must ask quality questions, develop new investigations, and explain those results in a manner all can understand without ambiguity. This process is known as inquiry. "Inquiry is the process scientists use to build an understanding of the natural world." (Networking for Leadership, Inquiry and Systemic Thinking, 2003) So, if inquiry is what scientists use to drive their own discoveries, our students in science education should be exposed to the same thought processes and experiences in the learning of science and how science works. There are many in the teaching profession who look at process skills and content knowledge as being a dichotomy. Inquiry learning brings the two together as a true definition of scientific knowledge and literacy. National Science Education Standards (NRC, 1996) states,

> "Scientific literacy is the knowledge and understanding of scientific concepts and processes required for personal decision making, participation in civic and cultural affairs, and economic productivity. It also includes specific types of abilities. (p.1)"

In stating this definition of scientific literacy, the National Science Education Standards call particular attention to processes and abilities as indicators of literacy. National Science Education Standards also define inquiry as,

> Inquiry also refers to the activities of students in which they develop knowledge and understanding of scientific ideas, as well as an understanding of how scientists study the natural world. (p. 2)

Without the use of inquiry learning in the science classroom, science becomes full of facts devoid of understanding (AAAS, 1989). Inquiry is the development of processes that allow students to investigate, acquire, and understand natural phenomena (NRC, 1996). These process skills are the link between science content and scientific understanding; between facts and knowing.

Inquiry and Hands-on Learning Interests Students

Students today are easily bored by instruction that was designed for classrooms in the 1950's. The students of today's world of technology, video games, and visual overload need their instruction to be interesting and active. Kanevsky and Keighley (2003) found students to be more interested in hands-on activities. They found that effective teachers realize that active learning is the opposite of boredom, and that active learning is actually the cure for boredom caused by antiquated methods of instruction. If students are bored by the content or presentation of their instruction, regardless of race, gender, ability, SES, or any other demographic that can be thought of, they will not perform. They must see a relationship to their everyday life. Students, who tend to be bored by school, do not tend to be bored away from school. This phenomenon indicates that they are not bored or boring people; rather they are not stimulated or challenged in school (Larson, 1989). Rhem (2003) stated that learning is most effective when students are actively engaged in creating knowledge and understanding by connecting what is being learned with their prior knowledge and experience. Kohn (1993) argued astutely that students need to take responsibility for their own behavior, but first we need to give them the ability to do so. He said that we should allow them a chance to live in a democracy today, not just learn about what it will be like when they grow up. That is to say, rather than being told what to learn, when to learn, and how to learn, students will take more interest in their own learning if they have input as to how the learning occurs and what is studied. In this way, students can take ownership of their learning. It is not something simply handed to them; it is constructed by them for them.

Tretter and Jones (2003) found that the use of inquiry learning caused an increase in the interest level of the science course and student participation, as well as an increase in academic performance. The range of achievement was less indicating that inquiry learning is better for all students regardless of ability level. They also found a decrease in absenteeism. Cawley, Foley, and Miller (2003) state hands-on inquiry does more than merely diversify instruction. They suggest that these activities offer teachers the opportunity for on-the-spot adjustments, allow students to raise and answer questions using different sources, enhance conceptual understanding by allowing for alternative representations, allow teachers to pace in order to account for different learning rates, and gives students the opportunity to demonstrate principles at high levels of generalization. Beaumont-Walters and Soyibo (2001) found that students from low socio-economic backgrounds and schools score lower when they are tested on scientific practices than those from a higher SES. They state the key must be a quality “hands-on, minds-on” curriculum that exposes all children to these scientific processes. In Georgia, as in most states, there is a rather large disparity between metro-Atlanta and rural Georgia in terms of tax base. There is simply more “local money” for the larger, more populated areas to use toward instruction. Students coming from low SES have fewer school-related experiences in their lives. Inquiry is a way to close the experiential gap that exists between areas of low and high SES.

Inquiry Learning in Science Classes Impacts All Students

Students With Disabilities

Students with Disabilities should be engaged with science concepts and inquiry learning. Science should be the easiest of content areas to mainstream SWD because they can use their own world to apply science concepts too (Atwood & Oldham, 1985). Teaching SWD in the science classroom is a difficult task for many teachers. Special education teachers tend to lack the science content to be effective just as the science teacher tends to lack the skills to adapt instruction to support SWD success (McCarthy, 2005). In fact, SWD receive less instruction in science than any other area (McCarthy, 2005). Because students with learning disabilities tend to have difficulty in reading and vocabulary, Gurganus and Schmitt argued that all science learning should begin with discovery rather than reading in order to allow students to formulate their own views of the natural world (Gurganus, 1995). SWD need additional supports, such as repetition and practice. Mastropieri and Scruggs (2006) found that differentiating instruction and the use of peer-tutors had significant effect on end-of-course examinations. SWD seem to thrive in science classrooms that have an environment of support, collaboration, and inquiry.

Economically Disadvantaged Students

Like SWD, Economically Disadvantaged (ED) students have had difficulty being successful in science classrooms as well. Minority and Economically Disadvantaged (ED) students have historically shown lower achievement in science courses. Literature shows a variety of issues ranging from stereotyping to lack of relevance in a student’s life (Lee, Buxton, Lewis, & LeRoy, 2006). These students tend to suffer from lack of science experience that often results from lack of resources, home influence, and a constant focus on literacy to the exclusion of other subjects (Lee, Buxton, Lewis, & LeRoy, 2006). Students who come from low socioeconomic areas and families tend to hold little interest for many traditionally taught science concepts. It is critical to engage these students in inquiry and project-based learning (Basu, & Barton, 2007). Basu et. al. (2007) and Lee et. al. (2006) found significant increases in achievement in minorities and ED students when engaged in inquiry and project-based learning.

Inquiry Learning – A History

"Students can learn about the world using inquiry. Although students rarely discover knowledge that is new to humankind, current research indicates that students engaged in inquiry build knowledge new to themselves."(NLIST, 2003) A student constructing and taking part in his/her own education is not a new instructional strategy; its roots lie in constructivist learning theory. Scientific inquiry is not necessarily the same as educational inquiry. Scientific inquiry usually results in new evidence that could lead to new discoveries or theories. Educational inquiry typically does not result in new information, but the construct for how a student will retain and apply his or her own new knowledge (NRC, 1996). The constructivist movement can be traced back to John Dewey and the progressive movement, Jean Piaget and Lev Vygotsky, and Jerome Bruner. Piaget may have coined the term "constructivist" when he referred to his views as being constructivist (Gruber & Voneche, 1977). John Dewey in an address to the American Association for the Advancement of Science said that science was more than just a body of knowledge; it included a process as well (Dewey, 1910). Joseph Schwab (1960) suggested that teachers take experiences from the laboratory and use them to build conceptual understandings in the classroom. The work of Schwab, Bruner, Dewey, and Piaget started a movement that placed at least as much emphasis on processes as it does on content. (NAS,

2000) The building of “knowledge new to themselves” is important because current brain-based research explains the need for students to internalize knowledge, thereby making it meaningful (Driver, et.al., 1994; Applefield, 2000). There are different interpretations of constructivism, but they tend to agree on four central characteristics believed to influence learning.

First, learners must construct their own learning. Learners must develop knowledge through an active construction process (von Secker, 2003; Tippins & Tobin, 1993). Much like a building requires a frame before true construction can occur, a student requires a frame, or scaffolding, on which to attach his/her learning. Content is learned differently by different students. Students place content into a comfortable placement within their own cognitive domain. Students place materials into their own “computer files” within their brain. In a sense, they arrange an internal, cognitive equilibrium. This cognitive equilibrium could be defined as self-confidence in acquired knowledge, which is very comfortable for students. Students feel a “balance” through that confidence that allows them to use that knowledge to make judgments and conclusions. However, in order for deep conceptual learning, or any difficult learning, to occur, this equilibrium must be upset in order to force the student to accommodate the new information and apply it in new situations.

Second, learning must be hinged on pre-existing knowledge and understanding while relating the new learning experiences to the experiences of the learner (Driver, et.al., 1994). Once a student has the scaffolding for his learning, true learning can only occur when a cognitive crisis occurs. A cognitive crisis places students in a position that requires them to reformulate or “re-scaffold” the new material or concepts into a workable, logical placement within the students’ memory. In other words, students realizes that the new knowledge does not conform to the old scaffolding, so students must reconfigure the cognitive structure to allow the new material to “fit.” The resulting disequilibrium requires the student to assimilate the knowledge into a new cognitive structure (Driver, 1989). The new structure results in a new understanding. Students must formulate the new knowledge into a cognitive framework that allows them to retrieve that knowledge and apply it to new situations. As stated earlier, as long as students are comfortable with their current state of understanding, there is no need for them to learn more. Without a student having to adapt, science or any subject can result in boredom or an over exaggerated view of one’s intellect and true conceptual understanding.

Third, there is a fundamental need for social interaction for deeper learning (Lee, & Paik, 2000; Driver, 1989). Just as scientists use a peer review process to justify findings, so must learners be allowed the chance to interact with classmates. Social interaction is the dialogue that confirms and clarifies knowledge acquisition (Driver, 1989). The learner must share findings and discuss data and observation in order to allow for the mental disequilibrium to subside. Just as scientists require peer review to confirm new discoveries and professional development, students require the same social interaction to discuss their new conceptual acquisitions. Process skills allow students to assemble and apply information and skills in a way that makes sense to students. Hands-on and inquiry-based learning techniques are core beliefs within the science education and scientific communities as shown through the AAAS *Benchmarks for Scientific Literacy* and the National Research Council’s *National Science Education Standards*. Both of these organizations believe inquiry learning through process skills is the key to student achievement and student mastery of science concepts.

The greatest improvement in student achievement is when students have the opportunity to construct their own knowledge through active involvement in the learning process itself (Yager, 1991). However, while inquiry-based learning does show an increase in overall science content mastery, it is sensitive to social contexts. Learning communities offer the opportunity for support and validation. Care must be given for the inclusion of all groups, with special concern given to integrate multiple level learners as well as different demographics. If these multiple levels are not understood, it could actually increase the achievement gaps (von Secker, 2002). Without the sharing of knowledge among peers, students who have not had access to experiences are still penalized by not being exposed to the experiences of those lucky enough to have them. There is a true deficit of knowledge based on a child's ability to attain it. There is also a wealth of knowledge often left untapped by not allowing students who have the knowledge to share their knowledge and experience. One of the major reasons for achievement gaps is the experiential differences among students (von Secker, 2002). The more experiences teachers can give students, even if those experiences are somewhat vicariously through other students, the greater the chances for achievement. It is important to point out this is not necessarily a case of the "haves versus the have-nots." All students bring experiences and knowledge into the classroom that can be an advantage in the classroom regardless of the extent of the experience. The varying perspective of students can allow for a deeper discussion and understanding of conceptual problems. The key is to allow students an environment that encourages and fosters sharing of ideas and use of inquiry methods. Students need to feel comfortable with their own understandings and know that those understandings can always be enhanced, strengthened, and even changed (Maroney, 2003).

Fourth, it is an absolute necessity for students to be given authentic learning tasks for meaningful learning. Meaningful student tasks that require students to apply knowledge to a situation where an understanding of science is required are the only way to fully assess a student's scientific literacy (Bybee, 2009).   That is to say, without true cognitive engagement on the part of the student, true assessment of the student's current state of understanding is not possible. Worksheets and tests generated by textbooks are not always good indicators of student understanding or achievement because of the lack of meaning within the student's life and the student's ability to apply and use the specified knowledge. Standards-based education relies on standards for all students. Therefore, students need to be cognitively engaged regardless of location or socio-economic class. Students need to show evidence of learning and mastery through the use of quality tasks that require the proof of that knowledge. Only through authentic cognitive engagement can students master important concepts, and only through the linkage provided through process skills can engagement and mastery be achieved.

For students to achieve these four aspects of the constructivist approach, teachers need to be prepared to support their students. Hodson (1996) summarized four steps that enable teachers to facilitate inquiry learning. First, teachers should identify students' ideas and views. In other words, know your audience. It will be important for teachers to recognize the academic and cultural backgrounds of their students. Understanding students' ideas and views allows the teacher to connect on even sensitive issues, such as evolution, without sacrificing important concepts or quality science. Second, teachers need to give students opportunities to explore their ideas. Let students have the opportunity to research and develop procedures to explore prior knowledge. Third, teachers need to stimulate students to develop, modify, and possibly change their ideas and views. Again, misconceptions must be thrown into conflict if students are to overcome them. Fourth, teachers need to support their students as they attempt to re-think or reconstruct their ideas. Teachers at all levels need to come to understand their roles as facilitators of knowledge.

Even at the collegiate level, research is showing loss of potential science teachers and science students because of perceptions and a lack of teaching for understanding vs. “I said it, you should understand it” (Seymour and Hewitt, 1997). Higher education has been slow to embrace inquiry learning in the natural sciences due to the belief in the traditional teaching methods. With the increased concern over science achievement and our position in the world with regard to science and technology, falling numbers of students interested in science are disconcerting, to say the least. Just as the kindergarten through twelfth-grade teachers have seen an increase in the emphasis of inquiry learning pedagogies; the higher education community is starting to feel that same emphasis. The National Science Foundation, through grants, are funding training for the natural science professors to learn new methods for teaching science in an attempt to keep more students in the field. That is not to say the standards are being lowered; standards are just being presented in a different fashion. This grant, known as the Partnership for Reform in the Instruction of Science and Mathematics (University System Board of Regents, 2003), is designed to enhance partnerships between thirteen school systems in Georgia and local universities. Kindergarten through twelfth grade (K-12) science and mathematics teachers work with professors in the natural science and mathematics departments, as well as those in the education departments, in order for each to learn from the other. In many instances, while the K-12 teachers gain content knowledge, the natural science and mathematics professors learn pedagogy, inquiry techniques in particular. The United States Department of Education, in partnership with the National Science Foundation, also gives each state a specified amount of grant money each year specifically designated to teacher quality. In order to be eligible for this grant (a total of 2.7 million dollars in 2004 and 4.4 million in 2005), school systems must be high needs schools as defined by poverty level or number of teachers teaching out of field, and they must have an active partnership with a university's science, mathematics, or engineering department. The purpose is to enhance content knowledge of fourth-grade through twelfth-grade science and mathematics teachers, but in working with these teachers, the professorial community also has the opportunity to learn pedagogies conducive to teaching content and the redelivery of that content to pre-college students (Georgia Department of Education Title IIA Competitive Grants, 2005).

Inquiry in Science Instruction and Assessment

The National Academies of Science in *Inquiry in the National Science Education Standards* (2000), suggest five essential features to inquiry. They are 1) engages in scientifically oriented questions; 2) gives priority to evidence in responding to questions; 3) formulates explanations from evidence; 4) connects explanations to scientific knowledge; and 5) communicates and justifies explanations. These features are completely meaningless without process skills. While these are the features that are included in quality inquiry learning exercises, students cannot engage in them without the proper tools. For instance, one cannot engage in scientifically oriented questions without proper skills for research. Students cannot formulate explanations from evidence if they cannot collect evidence and properly organize collected evidence. Scientific practices, as stated above, provide the link a student needs for deep understanding of science knowledge. Scientific practices are tools the science teacher must use to lead students toward inquiry learning just as a scientist uses them to explore and explain the natural world. Inquiry learning should be the goal of science educators, but this cannot be accomplished without the students first understanding how to employ tools, such as graphing, graph interpretations, questioning, and analyzing data. Olson and Loucks-Horsley (2000) said,

> In the science content standards, the ‘abilities’ of inquiry are skills and procedural knowledge that all students should be able to use in "doing science"--designing and carrying out an investigation.
>
> The "understandings" of inquiry include ideas about science as a human process for constructing knowledge—that scientists use mathematics and technology, for example, or that they undertake different types of investigations to answer different types of questions. In the science teaching standards, inquiry teaching and learning strategies are recommended as especially effective for learning the "big ideas" or important concepts of science. (Olson and Loucks-Horsley, 2000)

Alparslon et al. (2003) found that scientific practices are good indicators of understanding science content. With inquiry being the focus in science education, scientific practices are the precursory skills students need to understand true inquiry. The ability to graph, measure, write clearly and coherently are examples of scientific practices that lead the student to the development of scientific inquiry. Science teachers should use scientific practices to allow students to indirectly scaffold their ability to learn through inquiry. NRC (1996) stipulates that hands-on instruction alone does not insure quality inquiry learning on the part of the students. There must be intellectual engagement within the context of the activity to construct proper intellectual scaffolding for the mastery of content. There could be a tendency for teachers to perceive the use of hands-on activities and scientific practices as inquiry learning. Actually, these are the means to the end. Using hands-on and process based approaches in the classroom should be done with the goal for inquiry learning in mind. The Networking for Leadership, Inquiry and Systemic Thinking (NLIST) team operationally defines inquiry as,

> Student inquiry is a multi-facetted activity that involves making observations; posing questions; examining multiple sources of information to see what is already known; planning investigations; reviewing what is already known in light of the student’s experimental evidence; using tools to gather, analyze and interpret data; proposing answers, explanations, and predictions; and communicating the results. Inquiry requires identification of assumptions, use of critical and logical thinking, and consideration of alternative explanations. (NLIST, 2003)

It is important to understand that while scientific practices are subsumed under the operational definition of inquiry learning, they serve no true purpose if not used in order to move toward inquiry learning as a goal. Douglas Llewellyn defines inquiry by saying,

> Inquiry is the science, art, and spirit of imagination. It is the active exploration by which we use critical, logical, and creative thinking skills to raise and engage in questions. (Llewellyn, 2002)

Active exploration, creative thinking skills, and engaging students in questions are not skills students naturally acquire. They can develop naturally, but they need guidance and support. The use of tools, questioning, interpreting data, and predictions are all skills that students must attain in order to understand the full scope of learning science. At the same time, the use of these skills in the absence of rich content is also futile. Students will not make the connection between skills and content if just one or the other is presented.

In the age of high stakes testing, all strands of science must be assessed if students are expected to learn them. This means that inquiry learning must be assessed as well. Due to budget and time constraints, most states are not able to offer a true performance assessment as their statewide summative assessment where inquiry can be truly measured, so they must utilize other options to assess inquiry in order to attain information on the use of skills learned in the classroom. Interpretation of graphs, measuring using pictures of laboratory equipment, and formulating conclusions based on data are items that are easily assessed.

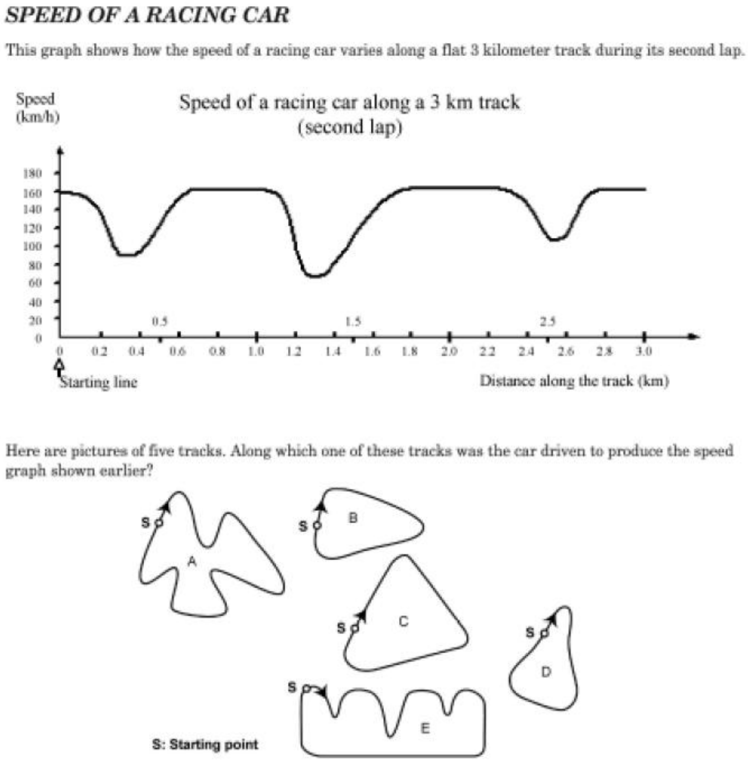
So what role can inquiry play in large scale assessments? Can it be assessed at all? Because science is perceived to be very content-rich, it is easy to see why the focus is on factual science knowledge. However, to fully engage a student into meaningful science assessment, all areas of knowledge must be accessed. One way to do this is to understand that real evidence of scientific understanding comes from the ability to integrate knowledge and skills and then apply those to new situations or address uncommon tasks (Bransford, Brown, and Cocking, 2000b). Problem solving utilizing scientific practice is a difficult, yet necessary feature within science assessment. The issue of statewide science assessment has different viewpoints. In a recent informal survey of the Council of State Science Supervisors, statewide science assessments were generally approved of (90%); however, there was more concern by several Council members as to how standardized tests would eliminate the use of inquiry in the classroom. It is possible to identify tests that contain elements of scientific practice and of problem solving skills. The construction of the tests must begin with this end goal in mind (Wilson & Bertenthal, 2005). In *Systems for State Science Assessment*, the committee identified five specific areas that can be assessed through a paper and pencil test. They are:

> identifying questions that can be answered through scientific investigations; developing descriptions, explanations, predictions, and models using evidence; thinking critically and logically to link evidence and explanations; recognizing and analyzing alternative explanations and models; and communicating and defending a scientific argument. (Wilson & Bertenthal, 2005)

These are key features that can be assessed in a large scale assessment provided that assessment developers are given explicit guidance as they develop items.

There are examples of assessments, both national and international, that have a deliberate focus on problem solving and use of contextualized scientific practices. Our own federally funded assessment, the National Assessment of Educational Progress (NAEP), implemented a new science assessment based on its new science framework in 2009. The framework was widely vetted, including focus groups conducted by the Council of State Science Supervisors (CSSS). This group represents individuals charged by their respective state departments of education to supervise science education within their state. Results from across the country and the resulting report to the National Assessment Governing Board (NAGB) showed the clear preference to use scientific practice and problem solving within the framework utilizing the context of the content (CSSS Focus Group Reporting Session, 2005). In other tests, such as Trends in Mathematics and Science Study (TIMSS) and in Program for International Student Assessment (PISA), the designers of the studies specifically target science practice in the context of content knowledge for the purpose of determining problem solving skills. TIMSS measures trends in the performance of students in grades four and eight in school mathematics and science in participating countries. PISA is administered to 15-year-old students and is designed to assess mastery of processes, understanding of concepts, and the ability to function in various situations. PISA assesses the areas of reading, mathematics, and science. Both studies are administered in the United States by the National Center for Education Statistics (NCES). A study conducted by Dossey et al. in 2006 showed that the two studies had different goals even though both wanted to evaluate student achievement. TIMSS focused more on what students should learn in school, while PISA focused more on the pure application of scientific knowledge in "real world" situations. The point of Dossey's study was to analyze the number of items that required problem solving skills in the comparisons of TIMSS and PISA. In order to accomplish this comparison, an operational definition was needed. An item was determined to be one that required these skills if "1) the context allows students to be engaged, 2) students do not have a known strategy to immediately apply, and 3) the situation calls for a solution" (Dossey et. al. (2006). Again, these are elements found within the auspices of scientific practice. Interestingly the result of the study showed PISA to be significantly higher in requiring the interpretation of information from a reading passage, while TIMSS had a higher number of problems requiring students to identify variables and relationships. Again at the onset, both assessments placed value in assessing scientific practices as viewed through content. Another study done by

Neidorf, Binkley, and Stephens in 2006 showed an even more robust analysis of NAEP versus TIMSS. Again, agreeing that both had a focus on scientific inquiry, NAEP had much more emphasis on inquiry and scientific practice than TIMSS did. In addition, TIMSS was found to have more than half of its multiple choice items as factual knowledge, whereas NAEP had more than sixty percent as conceptual understanding. In both assessments, use of scientific practice was prevalent in the construct of the test. Therefore, to increase the rigor and true assessment of science, scientific practice must be an integral part of large scale assessment from the construction of the assessment. An illustration of this type of item can be found throughout PISA, TIMSS, and NAEP released items. One such example is shown below.



SPEED OF A RACING CAR

This graph shows how the speed of a racing car varies along a flat 3 kilometer track during its second lap.

Here are pictures of five tracks. Along which one of these tracks was the car driven to produce the speed graph shown earlier?

In this example, not only does the student need to understand motion as it relates to velocity and acceleration, students also have to be able to interpret the graph accordingly. It would not be enough for them to just know how to interpret a graph to understand acceleration. This is what contextualized scientific practice is all about.

Do Scientific Practices Make a Difference?

Do the tools of inquiry learning have an effect on science achievement? There are two major trends in science education, inquiry-based learning and direct instruction. A 1983 study found that students exposed to a scientific practices/inquiry-based curricula showed greater achievement than students exposed to a traditional curricula based on facts, laws, and theories (Shymansky, Kyle & Alport, 1983). While there is a very real concern over accountability through state or national tests, there are studies that show inquiry learning and hands-on instruction can better prepare students for standardized tests (Stohr-Hunt, 1996). Alparslan et. al. (2003), found that utilizing science processes made a significant contribution to understanding respiration concepts. They stated in their study that the use of scientific processes helped students reshape and revise their prior knowledge and struggle with their misconceptions. By struggling with misconceptions, students were forced to realign their thinking with more appropriate and accurate knowledge. Alparslan also found that a student's ability to engage in scientific practice is a strong predictor for the understanding of respiration. Scientific practices are a necessity to understanding science, but it is important to look at some of the specific skills that lead to inquiry learning.

The National Research Council states that "inquiry is a way of finding out that involves questioning, observation, investigation, and discovery." (NRC, 1996) These are integral scientific practices that need to be taught and enhanced through classroom instruction. In the days of Aristotle and Socrates, teaching was done through asking questions. Part of learning science through inquiry is learning to ask and attempt to answer quality questions. Every great discovery was made because someone asked why or how. Chin (2002) states, “Questioning lies at the heart of scientific inquiry and meaningful learning.” That is to say, without allowing students to question, meaningful learning and discovery cannot effectively occur. Students’ ability to question goes beyond the science classroom into their general ability to problem solve. Watts, Gould et al. (1997) proposed three categories of questioning that show the progress of students’ learning. First, consolidation questions confirm explanations and tend to consolidate understanding. Second, exploration questions seek to expand knowledge. Third, elaboration questions allow students to examine, reconcile, resolve conflicts, and test circumstances. The latter is what science educators would like all students to be able to do when they leave high school. Allowing students to generate their own questions stimulates their interest, curiosity, and it encourages students to think about relationships among questions, tests, evidence, and conclusions (Chin, 2002). The results of Chin’s study showed that students ask better questions, showing greater understanding and achievement, when asked to utilize this process during an investigation rather than simply following the directions written by a book or teacher.

By having students design their own investigation, they are cognitively engaged throughout the entire process. When directions are always given, students may have a tendency to blindly follow directions. They may get answers, but no real knowledge to transfer to new situations. There is not true cognitive engagement. For students to design, they must process materials, procedures and data they receive into a workable form that students can use as well as structure it so other students can use their same procedure as well. Allowing students to design experiments gives the teacher proof of whether or not the students understands the goals of the activity and their ability to reason and think critically. Within the design of an experiment, students may have to troubleshoot problems with procedures or odd results. This is an opportunity not afforded to students who simply fill in the blank.

Effective lessons are found to include strategies that allow students to have a variety of experiences that enable students to employ multiple pathways in the development of conceptual understanding. Students' learning to process information and question their surroundings is the key to conceptual understanding. This is not a process that can be learned by utilizing a textbook; rather students must experience science. Bredderman (1982) found students whose teachers employ hands-on or inquiry learning achieved at a higher level on standardized tests than those who used the text. As a result of participating in inquiries, learners will increase their understanding of the science subject matter investigated, gain an understanding of how scientists study the natural world, develop the ability to conduct investigations, and develop the habits of mind associated with science.

Teacher Development for Inquiry Teaching

Pedagogical reform and teacher training using inquiry-oriented activities, while useful and sometimes fun, fall short of the true gains associated with inquiry learning. Klum and Stuessy (1992) suggest, "changes in curriculum goals (also) require concurrent changes in approaches used by teachers in improving learning." In the early days of state curricula, learning goals were very broad and contained little specificity. The curriculum must be viable and useful for teachers as well as be a tool to guide instruction. However, quality content is not enough. Quality content is usually present in most schools and districts. Weiss and Pasley (2004) showed eighty-nine percent of lessons showed significant worthwhile content. So, if quality content is present, there must be other factors that lead to student achievement.

According to Kanevsky and Keighley (2003), there are five factors that contribute to an effective classroom. The first is allowing students to feel they are in control of their own learning; the second is choice which allows students the ability to act on a chosen pedagogy; the third is that a class, or teacher, must present a challenge to the learner; the fourth is the complexity of environment and; the fifth is the caring attitude of the teacher. Students must take ownership in their own learning. When students feel empowered through ownership, they tend to show more effort at the mastery of concepts. Choice of pedagogy allows the student to accentuate his own learning by choosing the pedagogy most suitable for him or her. Challenge must be a critical issue in the classroom instruction. If the student is not challenged, there will be no cognitive disequilibrium, and therefore no learning. The use of inquiry learning and process skills automatically presents students with challenges by the very nature of its activities. The complexity of environment refers to the dynamics of the classroom environment. If the classroom is too simple, the learner may not feel the need to rise to the challenge; if too complex, the student may refuse to perform due to fear of failure. As much as anything, if the teacher does not care about his or her students, the students will not perform for the teacher.

Students are the beneficiaries of quality inquiry learning activities. It is important to use the term “quality”. For activities to take on the quality required to make a real difference in a student’s learning experience, the teacher must make good use of quality instructional decision making. Teachers who have been in the field for several years were probably taught using traditional teaching techniques. Training is needed to show teachers how to implement inquiry learning through the use of process skills in contemporary classrooms. Current science educational research indicates the benefits of inquiry learning for students. It is just as clear as to what teachers need to do in preparation for inquiry instruction (von Secker, 2003). The need for ongoing professional development is paramount. For most teachers, this is a completely new way of teaching. Olson S. and Loucks-Horsley point out,

> To teach their students science through inquiry, teachers need to understand the important content ideas in science -- as outlined, for example, in the *Standards.* They need to know how the facts, principles, laws, and formulas that they have learned in their own science courses are subsumed by and linked to those important ideas. They also need to know the evidence for the content they teach -- how we know what we know. In addition, they need to learn the "process" of science: what scientific inquiry is and how to do it. (Olson S. and Loucks-Horsley, 2002)

As more pre-service teachers enter the ranks of education professionals, the hope is that the science education paradigm will continue to shift toward the emphasis of processes in science rather than the factoids we tend to encourage. College and university education departments have embraced inquiry learning and use it in their science methods courses. The natural sciences have been more reluctant to begin this transition. It is a difficult transition to make because most of us were not taught in this manner.

In Georgia, an organization called the Partnership for Reform in Science and Mathematics or PRISM, has partnered universities with K-12 systems throughout Georgia in an attempt to enhance teacher training and student achievement (University System Board of Regents, 2003). The higher education community is planning to employ inquiry teaching/learning strategies in order to keep potential scientists and mathematicians in the programs. They will be trained in inquiry learning and best practices teaching strategies. All the while, the universities will aid in teacher professional development for the teachers in the systems surrounding their university. This partnership shows a commitment to improve science and mathematics instruction from both ends of public education. Ongoing relationships like this are a necessity to continue the shift toward quality science education.

Professional development for use of inquiry science in the classroom by teachers is the lynchpin to furthering science education. Huffman and Thomas (2003) conducted a study in which four types of professional development strategies were used. They are: 1) Immersion, to "do" science with a scientist or mathematician; 2) curriculum implementation, teachers using and refining the use of instructional materials; 3) curriculum development, teachers create new materials; 4) examining practice, examines through discussion the real world classroom; and 5) collaborative work, use of study groups, peer coaching, and mentoring. They found that curriculum development and examining practice were most related to standards-based instructional strategies. These two strategies were the best predictors of the future use of standards-based curriculum in the classroom. For true scientific understanding, students need to be involved and discussing observations; teachers need to be engaged in the same type of activity in order to best serve their students. Ball and Cohen (1999) state that professional learning should be a long-term, ongoing active engagement that allows for connections between the teachers' work and their students' learning. They should have the opportunity to practice and apply their new found knowledge in real world situations by experiencing job-embedded professional development (Peressini, Borko, Romagnano, Knuth, & Willis, 2004). "The emphasis is on a continuous cycle of exploring new issues and problems, creating cognitive dissonance, engaging in collaborative discussions, constructing new understanding, and improving professional practice." (Huffman et al. 2003) It is interesting that both teachers and students have the same needs with regard to the deeper conceptual learning that can make a difference in achievement as students and teachers.

Teacher Leadership through Mentoring

It was the opinion of the GaDOE that if changes were going to be made at a classroom level, intervention had to occur in those classrooms. The determination was made to utilize excellent practicing teachers to act as mentors or coaches to teachers in low performing districts. Much of the research on the topic of mentoring is focused on induction programs of new teachers in traditional and non-traditional certification routes. This program focused on supporting low performing schools and their teachers. What are mentors? How do they differ from principals or department chairs? All probably see themselves as mentors, but the specific duties and responsibilities of a mentor are different. Mentors mean different things to different people. One perspective is,

> people with career experience willing to share their knowledge; supporters, people who give emotional and moral encouragement; tutors, people who give specific feedback on one's performance; masters, in the sense of employers to whom one is apprenticed; sponsors, sources of information about and aid in obtaining opportunities; models, of identity, of the kind of person one should be to be an academic (Zeldtrich, 1990).

Another perspective is that they are coaches and to some who may not want them, nuisances. The idea of mentoring has been around a long time. The National Education Association (NEA) reports that mentoring programs have been in existence for approximately 50 years (Kent, 2009). Again, most of the time mentoring is discussed with entry into a profession, such as internships or induction, or even residency. NEA states that half the country now requires mentoring in some fashion for new teachers (Kent, 2009). Even given that, it is questionable how effective these programs are. Education culture tends to believe that once licensure is acquired, a person is completely ready to be on his/her own. Teachers are fully embraced as part of the profession without need of further interaction or oversight other than what is provided to veteran teachers (Danielson, 1996). There is evidence that mentoring not only retains teachers, but also improves achievement. In 2008, the National Education Association showed statistics that approximately 20 percent of all new teachers will leave the classroom within three years. In urban districts, it could be close to 50 percent (Kent, 2009). More alarming still, in 2007, the National Center for Educational Statistics suggested that a high percentage of teachers who leave the profession were under-supported and overwhelmed and this lack of support led to their departure (Kent, 2009). Mentoring is a key factor in implementing inquiry learning in the classroom. Melville & Bartley (2010) suggested that successful mentoring relationships have two parts: 1) the mentor must build an environment that allows the protégé to be self-sufficient and 2) the mentor must build an environment that leads toward teaching science as inquiry. That is to say, the protégé feels comfortable even if he/she must ask for help or make mistakes. In addition, the study revealed that a mentoring relationship must be based within a larger inquiry-based community if they are going to see real change in their practices (Melville & Bartley 2010).

Many mentor programs fail as a result of poor planning and no models or standards by which the programs are evaluated (Kent, 2009). At the inception of the Science Mentor Program (SMP), several items were developed based on current practice. Extensive training was put into place for the mentors, a list of protocols that govern contact with the school and system, a formula to determine which schools were eligible, and action plans to aid in the development of a plan to turn around the school (GaDOE, 2005). Another key factor in success is the preparation of mentors to be supportive and positive in moving change within the school (Saurino, 1999). It was also important to ensure accessibility to the mentors, so housing all of them in Atlanta was not feasible. Saurino found that a key feature to a successful mentor program was the ability to have ongoing contact with the mentor (1999).

Summary

Science education is still as complex as it was in the days of Sputnik. The research is clear about the best ways to present and teach science. It is clear about how to help students "think" like a scientist. What is also clear is that the content knowledge, pedagogical knowledge, and comfort of teachers are still not where these need to be to make real change in science education. Professors, science supervisors, science teachers, and even Presidents acknowledge the need of teaching quality science. The research tells us that we have to drill down into the world of the teacher, especially those who are struggling or those teaching in low performing schools. Teachers need help with inquiry and the implementation of inquiry-based lessons in the classroom. They need resources to help students learn. They sometimes need help with the actual content. One way to provide those things is through mentorship.

While the literature focuses on the mentoring of new teachers, it is clear that a focused mentoring program helps with the retention and performance of teachers. It is also clear from the research that more research is needed to generalize the work done by mentors to struggling teachers beyond induction into the workforce.

Chapter III. Methodology

Purpose of the Study

This study is designed to measure the impact of the Science Mentor Program (SMP) on student achievement as defined by the scores on the Georgia High School Graduation Test (GHSGT). Until 2005, all science achievement and/or science instruction improvements were conducted at the local school system or building level. The science portion of the GHSGT is required to receive a high school diploma in the state of Georgia. This was the first real attempt on the part of the Georgia Department of Education (GaDOE) to have an impact on science at the building level. The study employed a quantitative analysis of the science portion of the GHSGT for the years 2004 through 2007 to determine if the intervention by the SMP had significant effect on the science achievement within the cohort of schools, as compared to a set of schools receiving no intervention, on various subgroups within the schools, and on various levels of intervention within the SMP.

Research Questions

The research questions guiding this study are as follows:

1) Is there a significant difference in science achievement and proficiency for schools supported by the SMP each year in performance, between SMP and comparable schools, and between SMP schools receiving medium- versus high-level intervention on the science portion of the GHSGT from 2005-2007 during the period of intervention?

Null Hypothesis: Schools supported by SMP did not see significant differences in science achievement or proficiency from year to year and between SMP and comparable schools or between SMP high-level versus medium-level interventions as measured by the GHSGT in science.

2) Is there significant improvement in the science achievement and proficiency on the science portion of the GHSGT for subgroups (male, female; Economically Disadvantaged (ED), non-Economically Disadvantaged; students with disabilities (SWD), non-SWD; White, Black, Hispanic, Asian) within schools receiving high-level intervention by the SMP and between SMP and comparable schools from 2005-2007?

Null Hypothesis: The SMP does not result in a statistically significant improvement in science achievement or proficiency between subgroups within SMP supported schools or between SMP and comparable schools or high-level versus medium-level schools during the years of intervention as measured by the GHSGT in science.

Context of Schools

This study takes place throughout the state of Georgia. In 2004, when state leadership and the Georgia General Assembly decided that science achievement had reached crisis levels, a systemic model to improve science achievement was devised. With only 17 staff members to attempt to make and sustain change, the GaDOE decided to start with the schools in most need immediately. There was such need in the state, and so many systems asking for help, a fair and reliable system for school selection had to be developed. One of the Science Implementation Specialists (SIS), Juan Carlos Aguilar, Ph.D., developed and proposed a formula from which all the SMP schools would be selected. The formula, known to the SMP as the Culaca Formula, was vetted through the agency and adopted. The SMP design requires each SIS serve high-levels of support for five schools per year. These schools received high-level interventions meaning they received services of at least once per week. Once a school had 70 percent of first-time test takers proficient on the GHSGT in science for two consecutive years, they were completely removed from the active list, although SIS still stay in contact and provide support if requested and their scheduled allowed. In addition, other schools with high need, as determined by the Culaca Formula, were designated as medium-level intervention schools if the school system requested service and the SIS had room in the schedule to support the school at least two times per month. Highest priority went to schools identified as high-level intervention due to the fact the purpose of the program was to turn science achievement around one school at a time.

The Culaca Formula is based on the outcome of a need factor, nf. The overall nf is calculated as a weighted average of partial need factors calculated in GHSGT, End-of-Course Tests (EOCT), Adequate Yearly Progress Status (AYP), graduation rate, and number of students. Due to the fact the program is focused on improving science performance on the GHSGT and thereby the graduation rate, the GHSGT accounts for 25% of the overall nf. EOCT, AYP, graduation rate, and number of students account for 20% for each EOCT, 15%, 15%, and 5%, respectively.

Partial nfs are calculated by listing the schools from lowest to highest. The distribution of values is divided into quartiles and each quartile listing is divided in half. Each category is assigned a value between 0 and 8 with 8 representing the highest need. The exceptions to the nf category calculation were the partial nf for AYP determination and student population. Because so much intervention already takes place in schools who have been on the AYP Needs Improvement list (NI) and is required by *No Child Left Behind* (NCLB), the GaDOE decided to focus attention on schools at the upper end of the NI scale (NI 6-8) and schools about to enter contract monitored status in NI 3-4. Student population was viewed greatest to least because of the priority to help as many students as possible early in the program. Once a partial need factor is established for each category and school, the overall nf is calculated according to the formula:

$$\text{Overall Need Factor} = (.25)\ (\text{GHSGT } nf) + (.20)\ (\text{Biology EOCT } nf + \text{Physical Science } nf) + (.15)\ (\text{AYP } nf + \text{Graduation Rate } nf) + (.05)\ (\#\text{ students } nf)$$

where *nf* represents the partial need factor.

Once each school was assigned an overall nf, each SIS was assigned five schools as their priority schools (high-level intervention). For the most part, these were the schools actually serviced, however, the school system did have the right to refuse service and some did. For the purposes of this study, schools represented in the study accepted the services provided by the SMP. In the first year, 112 schools received services from the Science Mentor program. All but five schools identified as high-level intervention accepted support. GaDOE protocol requires permission from the superintendent and principal to receive the service. The five schools served under the same superintendent and SIS were not allowed to discuss the option with the principal. Since science is not a requirement under *No Child Left Behind* and because the SMP needed willing participants, these schools were not serviced. The GaDOE does not force non-federally required service on school systems or schools. The school system that refused services has had a history of refusing services from the GaDOE, so this was not completely unexpected. For the purpose of this study, only schools who received consistent levels of intervention (high- or medium-level) from 2005 through 2007 were used in the study reducing the total number of schools available for this study to 71, 49 high-level and 22 medium-level. The focus of the study is on the 49 high-level intervention schools. However, an analysis will be performed on the comparison between schools receiving high- and medium-level interventions.

Description of Intervention

The focus of this study is on schools that were given medium- to high-level intervention. A school identified for medium-level support received on-site service two times a month at a minimum. High-level support received on-site service at least weekly. During on site visits, SIS would perform several required tasks and additional tasks as needed by the school or teachers. Schools in this study received the same level of intervention for school years 2005 through 2007. All schools developed Action Plans with the help of the SIS to monitor progress.

Required tasks included a pre-visit interview with the Department Chair (DC). During this meeting, they would first discuss the progress of the staff toward the Action Plan since their last visit. The SIS would share plans for the day with the DC regarding the work flow for the day and any potential challenges or concerns. The SIS would also collect any feedback from the DC that would be helpful in providing service. The SIS is required to meet with each teacher each visit to discuss support. This can be done as a group or individuals during planning, lunch, or before/after school. The focus is on teachers who deliver instruction in Biology and Physical Science. During these meetings, the SIS and the teachers plan for the next visit, review the day, or discuss the results of jointly planned activities. An exit interview with all staff is also a requirement for the purposes of debriefing the day and agreeing on next steps for the next visit. The SIS always meets briefly with the principal upon availability. It is important to point out that the SIS does not report an “evaluation” to the principal. As mentors, the SIS have to gain the confidence and trust of the teachers with whom they work. This would not be possible if the teachers felt they were being evaluated.

Additional tasks include mentoring new teachers, planning for instruction, modeling lessons, critiquing instruction, and providing additional resources if the teacher does not have access, knowledge or time. These activities were shared across the SMP. Because each SIS had their own level of expertise, some in biology, others in chemistry, they would constantly share experiences and products developed over the GaDOE listserv and staff meetings. This way, best practices were shared statewide and any teacher in the identified schools had access to a large number of resources. Many times, SISs found teachers needed the most help in finding the resources and implementing the inquiry instruction needed to prepare students for scientific literacy and the GHSGT (Aguilar, 2009).

Research Design

The objective of this study is to evaluate the overall effectiveness of the SMP using a quantitative, quasi-experimental research design. The study utilized a quantitative approach for the purpose of determining if there was a statistically significant improvement in science achievement in schools that were supported by the SMP from year to year, SMP supported schools as opposed to a similar control group, and between different levels of intervention within the SMP. It is a quasi-experimental study as neither the treatment schools nor the control schools were randomly assigned. A mixed ANOVA was selected for each question as it allows each GHSGT administration to act as its own control group from year to year to add to the robustness of the evaluation. The year-to-year average scale score or proficiency rate for the GHSGT in science was used as the dependent variable. The study utilizes a 3 (Treatment Groups: SMP high-level intervention schools, medium-level intervention schools, non-SMP schools) x 4 (GHSGT school years: 2004, 2005, 2006, 2007) analysis of variance for question one. The treatment on the school represents the between effect while the scale score or proficiency rate from year to year represents the within effect. For question two, mixed ANOVAs were used to compare the identified subgroups. Gender, economically disadvantaged/non-economically disadvantaged, and Students with Disabilities/regular education students will be analyzed using a 3 (Treatment Groups: SMP high-level intervention schools, medium-level intervention schools, non-SMP schools) x 2 (Comparison groups) x 4 (GHSGT school years: 2004, 2005, 2006, 2007) analysis of variance. A 3 (Treatment Groups: SMP high-level intervention schools, medium-level intervention schools, non-SMP schools) x 4 (Ethnicity: white, black, Hispanic, Asian) x 4 (GHSGT school years: 2004, 2005, 2006, 2007) analysis of variance will be used to analyze gender. The subgroups within the three school groups are between effects while the scale score or proficiency rate from year to year represents the within effect.

Sample and Setting

The study uses school-based GHSGT data. School-level is the most accurate analysis due to the fact the SMP conducted services at the school-level. The study involved a total of 71 schools identified in the original cohort as medium- or high-level interventions by the SMP for a minimum of two years. Data from the science portion of the GHSGT between 2005 and 2007 was used exclusively. In 2008, the GHSGT was revised with a new scale score and cut scores. In years previous to 2008, the GaDOE statistically equated the GHSGT from year to year in order to maintain validity and reliability. The study was conducted using data from first-time test takers only. Students who were retaking the test were not included in the sample. For the purposes of this study, three groups of schools were analyzed; schools receiving high-level intervention (Group A) from the SMP, schools receiving medium-level intervention from the SMP (Group B), and schools not receiving SMP intervention (Group C).

Group A represents the 49 schools in the first SMP cohort that were identified as requiring high-level interventions. In Group A, four schools are considered Urban, two are considered suburban, and the remaining 43 are considered rural. Group B represents the 22 schools identified as receiving medium-level intervention. In Group A, six schools are considered Urban, three are considered suburban, and thirteen are considered rural. Group C represents 49 schools that are representative of Group A with regard to subgroup percentage (white, black; male, female), percentage of free and reduced lunch (FRL), students with disabilities (SWD), geographic location, and population who received little to no support from the SMP. The criteria for selection of Group C were as follows: 1) Percentages of white, black, FRL, and SWD; 2) geographic location; and 3) population. The researcher looked for ranges of five percent or less within the subgroup selection as a basis for similarity as the first factor in determining similarity. Most ranges were less than three percent for each subgroup. The preference was to select similar schools in the same general geographic location. When that was not possible, a school of similar size was selected from another area of the state. If a similar size was not available, the top two ethnic percentages and SWD percentages were used as the final determining factor. As a cross reference, the researcher used www.georgiaeducation.org to find the Similarity Index for each school. Georgia Education.org is a tool furnished by the Georgia School Council Institute for the purpose of identifying and comparing schools. Their data is taken directly from the GaDOE and the Governor’s Office of Student Achievement (GOSA). The Similarity Index is based on (1) Percentage of students eligible for free or reduced price meals (FRL); (2) Percentage of students with Limited English Proficiency (LEP); (3) Highest ethnic percentage at the selected school; and (4) Second highest ethnic percentage at the selected school; this renders a Similarity index on a scale of 1 through 6 with 1 being most similar. The Similarity Index can also be seen in Appendix C.

Methods and Procedures

GHSGT data for each of the schools in the study were attained from the Georgia Department of Education. School-based data is available through the GaDOE testing website. All data is free and considered to be in the public domain.

The within-effect is designed to specifically test if gains in achievement as measured by the schools' GHSGT average scale score or proficiency rate and their subgroup average scale scores or proficiency rate increase significantly. This is an advantage as a mixed ANOVA is generally more powerful as it is less likely to produce Type II errors (Huck, 2000) where the study could produce an effect that is not really present. This is a particular advantage as the groups are being analyzed at different periods of time as opposed to different treatments of the two groups. It must be noted that this type of analysis does come with potential limitations. The key assumption is that variances within the population exhibit similar patterns. Ad-hoc testing will be required if the F-value shows the assumption to be false.

The between-effect is designed to specifically test if the gains exhibited in SMP schools and their subgroups is significantly different than non-SMP schools and their subgroups. For purposes of analyzing the overall effectiveness of this program, it is important to evaluate changes in achievement in SMP schools against the changes of a control group to assess whether the treatment by the SMP was a contributing factor to increases in achievement for those schools. The same analysis is important to determine if the different levels of intervention are significantly different as policy decisions are made going forward.

GHSGT Instrument

The metric used to evaluate change is the scale score for the science portion of the GHSGT. Use of this as a metric, in addition to steps taken each year by the GaDOE, ensures comparison from year to year. Each year, once the tests are statistically equated, students are assigned a scale score based on the same range each year. This is done to ensure longitudinal reliability. While the test may show slight fluctuations in “hardness” each year, the scale score provides reliability between student scale score in 2005 versus a different student in 2006. A more in depth discussion of the validity and reliability of the GHSGT follows.

The study centers on the scale score of the GHSGT as opposed to the percent meeting or exceeding standards due to the fact that year-to-year analysis can best be studied using a vertical scale. Percent meeting or exceeding each year is not as useful a metric since different students take the test each year. Due to the internal reliability the GaDOE uses to ensure fair and valid results each year, scale score will be a better metric to measure significant change over a period of years.

GHSGT – A History

The Georgia High School Graduation Test was put into place as a result of Georgia state law (O.G.G.A., Section 20-2-281) requiring students who entered ninth grade after July 1, 1991 to pass curriculum-based achievement tests in order to receive a high school diploma. Development of the test began in 1991 with item specification and item bank development in the areas of English Language Arts (ELA), mathematics, science, and social studies. After three years of field testing and analyzing results, the first operational test was administered in 1994. Performance standards, or cut scores, were set for ELA and mathematics in 1994 with social studies following in 1996 and science in 1997 (GaDOE, 2006).

The GHSGT and supporting documents such as the Content Descriptors and Performance Descriptions were developed based on the Quality Core Curriculum (QCC). The original QCC was not developed to be tested. It was designed to give guidelines to teachers as to what they should teach. It was found to be the typical "mile wide, inch deep" intent by a Phi Delta Kappa report in 2002. As such, the assessment also struggled to get beyond rote memory type questions and the items were derived from a large variety of content. The GHSGT also had separate domains for science process skills and content. Items were developed without considering context meaning that students were tested on reading a graph, not on whether they could use a graph to answer science content questions. This was changed in 2005 with the implementation of the GPS. A chronology of the development of the GHSGT is attached in Appendix N.

GHSGT Development Process

The GHSGT is required for students to receive a high school diploma in the state of Georgia. As such a high-stakes test, the GaDOE has had to take deliberate steps to ensure the validity and reliability of its tests. A big component to the validity of the test is the process of test item and form development. The process is clearly explained in several documents on the GaDOE Testing website, http://www.gadoe.org/ci_testing.aspx?PageReq=CI_TESTING_GHSGT. One particular description comes from the 2007 GHSGT Technical Manual. This manual is published yearly and contains background, development processes, and data on validity and reliability. A summary of the process of development from the 2007 technical guide is below.

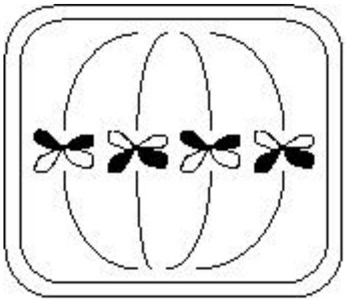
- *Identification of Test Content Domain: Committees review curricula, textbooks, and instructional content to develop appropriate test objectives and targets of instruction. Committees provide advice on test models and methods to align the test with instruction.*
- *Development of Test Specifications: Committees of content specialists develop test specifications that outline the requirements of the test, such as eligible test content, item types and formats, content limits, and cognitive levels for items. These specifications are published as a guide to the assessment program.*
- *Development of Items and Tasks: Using the test specifications, GaDOE staff and PEARSON work with the item development contractor to develop items and tasks.*
- *Item Content Review: All members of the contractor's assessment team review the developed items, discuss possible revisions, and make changes when necessary.*
- *Item Content Review Committee: Committees of educator experts review the newly developed items (some of which are revised during content review) for appropriate difficulty, grade-level specificity, and potential bias.*

- *Field Testing: Items are taken from the item content review committees, with or without modifications, and are field tested as part of the assessment program. Data regarding student performance, item difficulty, discrimination, reliability, and possible bias are compiled.*
- *Data Review: Committees of educators review the items in light of the field test data and make recommendations regarding the inclusion of the items into the available item pool.*
- *New Form Construction: Items are selected for the assessment according to test specifications. Selection is based on content requirements as well as statistical (equivalent passing rates and equivalent test form difficulty) and psychometric (reliability, validity, and fairness) considerations.*

A key feature to the process in Georgia is the level of involvement of classroom teachers at all levels of development. Georgia is one of a handful of states that has classroom teachers involved in all phases (Fincher, 2009) of test development. Many teachers express this to be one of the best professional development opportunities they have experienced that impacts their classroom practice (Fincher, 2009). Based on comments like this one, Science Mentors were exposed to this process and allowed to observe the entire process. This allowed them to enhance their own knowledge of large scale assessment and building quality assessments, but they were also better prepared to explain the process to the classroom teachers they support (Aguilar, 2009).

GHSGT Science Items

As stated earlier, science items on the QCC version of the GHSGT were very discrete. They tested one concept or skill and did not allow for the combining of concept and skill. An example of this is shown below with permission from the GaDOE.



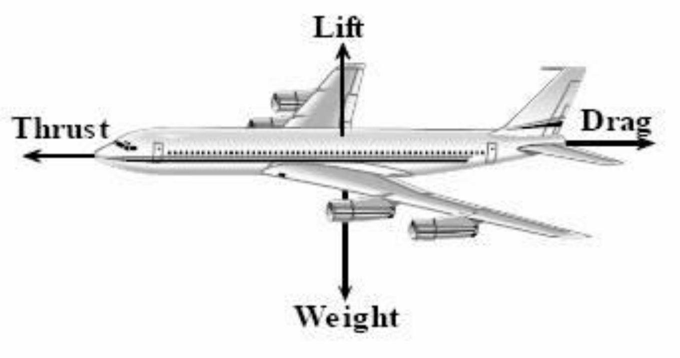
The plant cell shown above is in which phase of mitosis?

(A) anaphase
(B) interphase
(C) prophase
(D) metaphase

In this item, students are simply required to have memorized the phases of mitosis and pick the appropriate answer. These types of items are not necessarily meritless at the classroom level. Classroom teachers certainly need to know if students know basics before putting them into new or unique circumstances. As a matter of philosophy, the GaDOE and committees of teachers felt the GHSGT needed to match the GPS in intent as well as the written word. Therefore, the decision was made to "double-code" items going forward on the GHSGT. This decision required items to be developed using the content and scientific process standards. This also added to the complexity of the items taking many of them to a higher level of Webb's depth of knowledge (Fincher, 2009). Other states were used as models of this philosophy such as Virginia and Massachusetts. An example of a GHSGT science item that requires students to use Characteristics of Science and the content to answer the question is listed below. Additional Georgia items as well as items from Virginia and Massachusetts are shown in Appendix O.

GHSGT Example

An airplane in level flight is acted on by four basic forces. *Drag* is air resistance, *lift* is the upward force provided by the wings, *thrust* is the force provided by the airplane's engines, and *weight* is the downward force of gravity acting on the airplane.



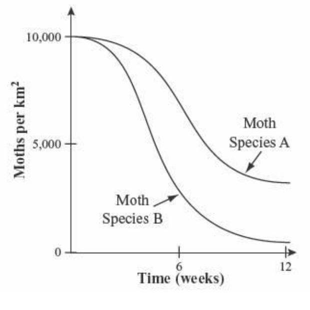
In level flight at constant speed, which pair of forces must be equal?

(A) lift and drag
(B) drag and weight
(C) lift and weight
(D) thrust and lift

For this problem, a student has to be able to understand how to read the diagram and understand how the arrows represented forces as well as have an understanding of balanced forces to answer the question.

Massachusetts Example (Released to the public, 2005)

The praying mantis is a predatory insect that often eats moths. The graph below shows the relative numbers of two species of moths over 12 weeks after the introduction of the predatory praying mantis.



What characteristic of this ecosystem is **best** indicated from this graph?

(A) Species B was preferred as food over species A.
(B) Species B may replace species A in this environment.
(C) Species B will reproduce more rapidly than species A.
(D) Species B was more abundant at the beginning of this time period than species A.

This item requires students to be able to interpret data from a graph and understand the relationship of organisms. The GHSGT has items of this type on the operational tests, but have not been released for public use as yet. Due to the transition of QCC to GPS, many items such as these have been kept by the GaDOE for use on their On-line Assessment System (OAS). The OAS allows teachers to make their own benchmark assessments, therefore these items are not for public view until the bank of items is such that the GaDOE can release them without harming the integrity of the bank.

GHSGT Validity and Reliability

The GHSGT and all tests developed for the state of Georgia go through a rigorous process to ensure their validity and reliability. The GaDOE is committed to using the *Standards for Educational and Psychological Testing* (1999) as developed by the American Educational Research Association (AERA), the American Psychological Association (APA), and the National Council on Measurement in Education (NCME). These standards were developed "to promote the sound and ethical use of tests and to provide a basis for evaluating the quality of testing practices" (SEPT, 1999). These standards hold critical the need to collect evidence in order to establish validity and reliability in all large scale assessments as well as guidelines as to what test developers should do to ensure a quality test. The GaDOE is committed to both the validity and reliability needed for high quality assessments. GaDOE asserts in its 2008 Validity and Reliability Brief, "While validity is the most important consideration in the test development process, a test cannot be valid without a high degree of reliability" (GaDOE, 2008).

Since the same students will not be taking multiple tests, it is important to discuss the validity and reliability of the GHSGT. Evidence of validity is collected in two main categories, content and construct validity. Content validity addresses the premise that a test measures what it claims to measure (GaDOE, 2006). The GHSGT is a criterion reference test, and as such, it is defined by the content is it suppose to measure; in this case, the GPS. Several pieces of evidence are used to support content validity. Each year, committees of teachers, test developers and GaDOE staff meet to discuss operational and field items to ensure their alignment to the curriculum. If relevance to the GPS is determined to exist, an item is passed along to the next phase of either testing or operational use. If the item is found to not have relevance, the committee may offer revision, in which case the item must be re-field tested if it has already been tested, or the committee may reject the item outright. In addition to educators and test developers having active roles, educators also validate the alignment by agreeing on the match between the item and the content standard it was conceived to measure. In addition, the tests are validated by using the testing blueprint to ensure the number of items on the assessment match the initial construct of the test as developed by educators.

Construct validity is a measure of the degree to which the test score is a measure of the psychological characteristic (i.e. construct) of interest (GaDOE, 2006). That is to say, the score must have the ability to generalize to the degree of what is attempting to be measured is actually present. If we are to test a construct such as predicting genetic variability, the test items must be able to allow for the generalization that accurate answers indicate a student understands genetic variability. The collection of construct validity evidence is a continuous process. Two metrics are used for the GHSGT, item point-biserial correlations and Rasch fit statistics. A point-biserial correlation is the correlation between how the performance on an item (percent of students who got the item correct) and the overall test score. A high point-biserial, 0.30 or above, indicates that students who performed well on the overall test would have also gotten the item correct and those students who did not do well overall, got the item incorrect. In addition, a high point-biserial acts as an excellent indicator in supporting the reliability of the assessment and its items. Another way to say this is that the item successfully discriminated between high and low performing students. Rasch item fit statistics show how well the items fit the evaluation model. Rasch is another method to determine difficulty of an item and how that item performs on the assessment.

Reliability is simply the ability to obtain consistent measures (GaDOE, 2006). The ability to obtain consistent measures is required to make appropriate interpretations of test scores. Reliability is based upon the premise that a “true score” exists without the observation of a measurement error. Measurement errors always exist; therefore an “observed score” is the result of a measurement. Mathematically, reliability can be defined as

$$Reliability = \frac{TrueScore.Variance}{ObservedScore.Variance} = \frac{TrueScore.Variance}{TrueScore.Variance + ErrorScore.Variance}.$$

If no error existed, the mathematical result of the ratio above would equal 1. Reliability should therefore be as close to 1 as possible. Cronbach’s alpha is used by the GaDOE to determine reliability of the GHSGT. Cronbach alphas calculated greater than 0.80 indicate acceptable reliability among test forms and subgroups. Table 1 shows the alpha scores on the 2004 through 2007 spring administrations of the GHSGT.

Table 1 Georgia High School Graduation Test Multi-Year Cronbach’s Alphas

	2004 Spring Administration	2005 Spring Administration	2006 Spring Administration	2007 Spring Administration
	Science	Science	Science	Science
Mean Cronbach’s alpha	0.92	0.92	0.92	0.92

In addition to Cronbach’s alpha, the traditional standard error of measurement (SEM) is calculated to be used with the estimate of reliability to make determinations as to the degree to which the measurement error is influencing individual scores. The SEM is based on the premise that items such as science achievement cannot be measured without a degree of error. SEM expresses unreliability in terms of the raw score metric. The SEM places an "error band" around the individual score that indicates how much the error may have affected the score. The SEM can be calculated using the following formula:

$$SEM = \sigma_x \sqrt{1 - \rho_{XX'}} \ ,$$

where, σ_x is the standard deviation of the total test (observed measure scores), and $\rho_{xx'}$ is the reliability estimate for the test. If it is assumed that the errors are normally distributed throughout the testing population, the correct score should fall within the error band approximately 68% of the time if the test was repeated multiple times. The SEMs for the spring administrations of the GHSGT are listed in Table 2.

Table 2 Georgia High School Graduation Test Multi-Year Standard Error of Measurement

	2004 Spring Administration	2005 Spring Administration	2006 Spring Administration	2007 Spring Administration
	Science	Science	Science	Science
SEM	3.42	3.69	3.54	5.24

GHSGT Equating

In order to ensure consistency of passing standards across different test administrations, GaDOE and its contractors construct all GHSGT to be of similar difficulty within the content areas. GaDOE uses a two stage statistical process with pre- and post-equating stages. Pre-equating utilizes item parameters from previous operational and field tests to construct a similar test. Each operational test uses embedded field items allowing for the linking between field test performance and operational performance. Table 3 reports the pre-equating values required to construct equivalent tests.

Table 3 Georgia High School Graduation Test Pre-Equating Values

	Rasch Difficulty		**Point Biserial**	
	Mean	**SD**	**Mean**	**SD**
Science	-0.423	0.693	0.332	0.080

Post-equating reviews the administration for inconsistency with previous operational tests. If items are found to have different outcomes than in previous use (field or operational) appropriate adjustments are made to difficulty estimates before the scale scores are computed, thereby allowing for fluctuations in "hardness" (GaDOE, 2006). GaDOE also uses "linking items" to scrutinize performance from year to year. These linking items are used year after year to view performance. Approximately 25-30% of the testing bank is carried forward from year to year. Equating takes place for the spring and fall administrations as they contain adequate representation of the testing population.

GHSGT Scale Score

The GaDOE derived its scale score system by using statistics gained after a test was equated to the baseline test form. The purpose of a scale score system is to report consistent information about student performance from year to year and administration to administration. Because each test form possesses a different degree of difficulty, forms and administrations must be equated to ensure scores are comparable. Therefore, once a passing score is established on the very first spring administration, the passing standard will always be the same. Currently the passing score on the GHSGT is 200 which was set when the first GPS ELA/Science GHSGT was administered. Therefore, all GHSGT will have a passing score of 200 regardless of the subject. This way, 200 will always imply the same level of student ability (GaDOE, 2007). The equating summaries for spring administrations from 2004 to 2007 are in Table 4.

Table 4 Georgia High School Graduation Test Equating Values Multi-Year Summary

	2004 Spring Administration	2005 Spring Administration	2006 Spring Administration	2007 Spring Administration
	Science	Science	Science	Science
# Items	70	69	70	70
# Students tested	98,537	82,256	79,062	92,454
% Passing	68	71	76	78
Form mean	48.19	47.53	47.82	52.54
Form SD	12.046	12.08	12.53	11.71
Cut score	47	46	42	47
KR-20 (Cronbach's Alpha)	0.92	0.92	0.92	0.92
SEM	3.42	3.69	3.54	5.24

GHSGT Proficiency Rate

A key metric in the evaluation of the Science Mentor Program is the proficiency rate within schools as defined by the percent of students meeting or exceeding the assessment standard. Each time there is a significant change to an assessment, the GaDOE resets the standards for its assessment. As with the process for ensuring validity described in previous sections, Georgia teachers set the minimum raw score a student needs to receive to be considered as meeting or exceeding the standard. The GaDOE utilizes an often used methodology called the modified Angoff. The modified Angoff is a process by which teachers set the minimum cutscores for meeting and exceeding the standard using their judgment on an actual test form, statistical performance of students on each item, and predicted outcomes of the test form. This process is also utilized by the National Assessment of Educational Progress and states such as Massachusetts and Virginia (Fincher, 2009). As discussed above, each year, tests are statistically equated to ensure validity and reliability across years. Therefore, the minimum raw score for students to meet standards may fluctuate slightly depending on the statistical difficulty of the test. In short, if a test form is statistically more difficult, a student may not have to have as many items correct as a student who takes a test form slightly less difficult. This ensures equity and validity since not all students take the same test form within and between years.

The proficiency measure is critical for this study as it was used to procure the funds for the SMP. This is the measure the map in Figure 1.1 is based upon. While student achievement as measured by the scale score is certainly important for this study, the proficiency rate is just as important in order to determine if the improvement in achievement translated into more students being proficient on the science portion of the GHSGT.

Chapter IV. Results and Analysis

Overview

In this study, the researcher analyzed the effect of the Science Mentor Program (SMP) on identified schools with regard to their science achievement and proficiency rate. Science achievement was defined as the average scale score for each school on the science portion of the Georgia High School Graduation Test (GHSGT). Proficiency rate was defined as the percent of first-time test takers who met or exceeded the minimum score (cutscore) on the science portion of the GHSGT. Students were required to meet or exceed a scale score of 500 to be considered proficient on the GHSGT. The study analyzed the year to year improvements in scale score and proficiency rates in SMP schools (two intervention levels: high- and medium-level) and compared those schools to a set of schools that received no official intervention to determine the effects on all students, gender, ethnicity, Economically Disadvantaged (ED), and Students With Disabilities (SWD). All analyses used mixed analyses of variance to determine within-effects and between-effects. The researcher utilized Predictive Analysis Software (PASW) Statistical Package for the Social Sciences (SPSS).

A key assumption to evaluate the validity of repeated or mixed ANOVA is the sphericity assumption. This assumption states that population variances associated with the levels of the repeated measures factor must exhibit at least one of a set of acceptable patterns. The test for this assumption is called Mauchley’s Test for Sphericity. If a Mauchley’s test is significant, sphericity is compromised. A common method to account for violation of sphericity is to use the Geisser-Greenhouse approach. The Geisser-Greenhouse's Epsilon was used for any analysis that violated the sphericity assumption. Use of this method allows the researcher to compute a more conservative F-test that overcompensates for sphericity violations (Huck, 2000). Pairwise and Scheffe's post-hoc tests were also used to confirm statistical differences in effects. For the purposes of reporting the analyses of this study, the levels of effect size were set according to Jacob Cohen's criteria. Effect sizes of .20, .50, and .80 are considered to be small, medium, and large, respectively (Huck, 2000). The same thresholds were used to determine the level of statistical power.

Analysis

Research Question 1

Is there a significant difference in science achievement and proficiency for schools supported by the SMP each year in performance, between SMP and comparable schools, and between SMP schools receiving medium- versus high-level intervention on the science portion of the GHSGT from 2004-2007 during the period of intervention?

All Students Analysis

The researcher performed a 3 x 4 mixed analysis of variance (ANOVA) for achievement (scale score) and proficiency (percent of students meeting or exceeding standards). In each analysis, the GHSGT Administrations (scale score and percent proficient, respectively) represented the within-subjects variable and the treatment level represented the between-subjects variable. In both analyses, only the performance from year to year showed significant gains. A summary of the findings from the mixed ANOVA appears in Table 5.

Table 5 All Students Analysis– Scale Score and Proficiency Mixed ANOVA Findings

	Scale Scores			Proficiency Scores		
	df	F	Partial Eta Squared	Df	F	Partial Eta Squared
Between Subjects						
Treatment Group	2	1.278	.017	2	.951	.022
Error	112			111		
Within Subjects						
Administrations	3	27.787***	.144	3	18.608***	.199
Administrations X Treatment Group	6	1.002	.579	6	.789	.018
Error	336			333		
*p<.05 ** p < .01 *** p < .001						

All Students Scale Score Analysis

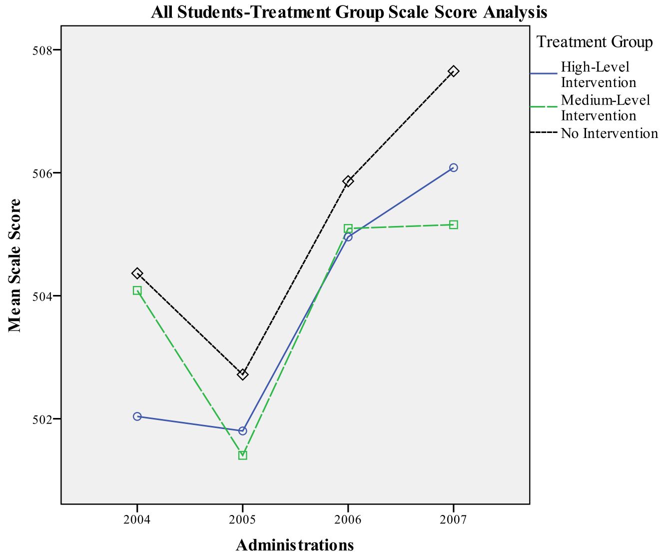
The first two-way mixed ANOVA, 3 (Treatment: High-level, medium-level, No intervention) x 4 (GHSGT Spring Administration Scale Score results: 2004, 2005, 2006, 2007), was designed to test the effects of the treatments on the average scale score of the school. Scale score was used as a measure of science achievement. During the time period assessed for this study, a student had to receive a scale score of 500 in order to be proficient on the GHSGT in science. Descriptive statistics for each variable are shown in Appendix D.

Mauchly's Test for Sphericity was found to be not significant (W=.961, p=.501) with acceptable Greenhouse-Geisser Epsilon values (.976). Therefore, sphericity can be maintained for the purposes of this study.

Tests of Within-Subjects Effects showed significant effects for each GHSGT Administration ($F_{3, 333}$=18.608, p<.001). The study showed a strong power (1.000) indicating the effects are without Type I error. While the study supported the increase in test scores from year to year, this was independent of the Treatment Groups ($F_{6, 333}$=.789). All three groups showed an increase in scores, but there was no significant effect due to high, medium, or no intervention. The 2004 administration was used as the baseline year for the study because it was prior to any intervention by the SMP. Each administration was significantly different when tested against the baseline year of 2004 ($p<.05$). Of particular interest is the most significant effect between 2004 and 2007 ($p<.001$). While the treatment groups did not appear to have significant effect within the administrations, the achievement as measured by scale score showed a definite increase since the program began. While the overall increase was significant between 2004 and 2007, pairwise comparisons showed significant interactions between 2005 and 2006 ($p<.001$). Changes between 2004 and 2005 ($p=.121$) and 2006 and 2007 ($p=.727$) were not significant.

Tests of Between-Subjects Effects confirmed a lack of significant difference between the means of the three Treatment Groups ($F_{2, 111}=.951$, $p=.389$). The power of this study is considered low (.211). While there was no significance found, it is important to recognize that in 2004, the high-intervention schools average scale score was 2.32 scale score points behind the no-intervention schools, but closed that gap in 2007 to only being behind 1.57 scale score points as seen in Figure 4.

Figure 4 – All Students Analyses Plot – Treatment Group Scale Score Analysis



While this is not a significant closure of the gap, progress was made in those schools. Medium-level intervention schools started virtually equal with the no intervention schools and actually did not show as much progress finishing in 2007 with a larger gap (2.5). In addition, the high-intervention schools started with a gap of nine scale score points behind the state average and finished in 2007 with a gap of eight scale score points. The no-intervention schools started seven scale score points behind the state average in 2004 and showed no gain by 2007 remaining seven scale score points behind the state average.

All Students Proficiency Analysis

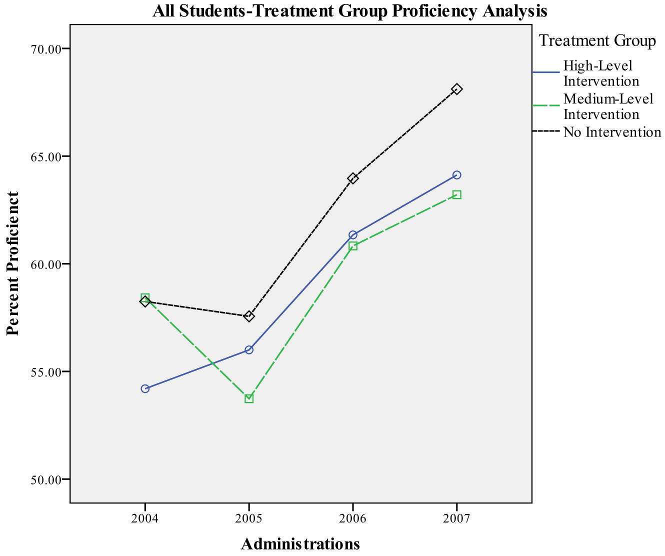
The second two-way mixed ANOVA, 3 (Treatment: High-level, medium-level, No intervention) x 4 (GHSGT Spring Administration Percent Proficient Results: 2004, 2005, 2006, 2007) was designed to test the effects of the treatments on the average percent of students meeting or exceeding the performance standard for the school. For the purposes of this study, percent of students meeting or exceeding the performance standard will be used to measure proficiency on the science portion of the GHSGT. As mentioned in Chapter 1, percent proficiency was used to bring attention to the crisis in Georgia's science instruction. Descriptive statistics for each administration and treatment group are shown in Appendix E.

Mauchly's Test for Sphericity showed a potential violation of sphericity ($W=.880$, $p<.05$). However, the Greenhouse-Geisser Epsilon value (.927) indicates that sphericity can be maintained for this study.

The Within-Subjects tests showed a significant increase ($F_{3,336}=27.787$, $p<.001$) in proficiency from year to year as shown in Table 5. While the study supported the increase in percent proficient for the treatment schools as shown in Figure 5, this was independent of the Treatment Groups ($F_{6,336}=1.002$, $p = .424$).

All three groups showed an increase in the percentage of students proficient, but none of the three groups showed a significant effect. As with scale scores, 2004 was used as the baseline year for the study because it was prior to any intervention by the SMP. 2004 was found to be significantly different than 2006 (p=.001) and 2007 (p<.001). The 2006 (p<.001) and 2007 (p<.001) administrations were significantly different than the baseline year of 2004. Pairwise comparisons were used to determine significant effects from year to year. Significant effects were found between the 2005 and 2006 (p<.001) as they were in the scale score analysis. There was no significance found between 2004 and 2005 (p=1.000) or 2006 and 2007 (p=.087) although 2006 and 2007 showed a substantial increase.

Figure 5 – All Students Analyses Plot – Treatment Group Proficiency Analysis



Tests of Between-Subjects Effects confirmed there was no significant difference in the percent of proficient students between treatment groups ($F_{2,112}$=1.278, p=.283). The power of the study and its effect size are considered small (Power=.272, Partial Eta Squared=.022). Pairwise comparisons and Scheffe's pot-hoc test confirmed no significant difference between the Treatment Groups. The high-level intervention schools started 4.05 percentage points behind the no-intervention schools in 2004 and slightly closed that gap through 2007 (3.99 percentage points) while the medium-level intervention schools saw their gap widen as with scale scores from virtually the same to a gap of 4.9 percentage points. With regard to the state comparison, high-level intervention and no-intervention schools maintained a 14 and 10, respectively, percentage point gap between 2004 and 2007. The medium-level intervention schools fell to a 15 percentage point gap.

These analysis of all students yielded data that showed that each of the three treatment groups showed increases over time, but those increases were independent of the treatment itself. Further analysis in Question 2 will provide a more detailed look at treatment's effect on the largest subgroups associated with these schools.

Research Question 2

Is there significant improvement in the science achievement and proficiency on the science portion of the GHSGT for subgroups (male, female; Economically Disadvantaged (ED), non-Economically Disadvantaged; students with disabilities (SWD), non-SWD; White, Black, Hispanic, Asian) within schools receiving high-level intervention by the SMP and between SMP and comparable schools from 2004-2007?

The researcher performed a mixed analysis of variance where the GHSGT Administrations represented the within-subjects variable and the Treatment level and Subgroups represented the between-subjects variables for both scale score and proficiency. The analyses of the subgroups were done separately to ensure clarity in the analysis. The analyses performed were as follows: 1) Gender (male vs. female); 2) Ethnicity (white vs. black vs. Hispanic vs. Asian); 3) Economically Disadvantaged vs. non-Economically Disadvantaged; and 4) Students with Disabilities vs. regular education students.

Gender Analysis

The researcher performed a mixed analysis of variance where the GHSGT Administrations represented the within-subjects variable and the Treatment level and gender represented the between-subjects variables for both scale score and proficiency. For both analyses, the only significant findings were in year to year performance and differences in performance of males versus females. These results are summarized in Table 6.

Table 6 Gender Analysis– Scale Score and Proficiency Mixed ANOVA Findings

	Scale Scores			Proficiency Scores		
	Df	F	Partial Eta Squared	df	F	Partial Eta Squared
Between Subjects						
Treatment Group	2	1.914	.017	2	2.441	.021
Gender	1	30.004***	.118	1	25.938***	.104
Gender X Treatment Group	2	.038	<.001	2	.056	<.001
Error	224			224		
Within Subjects						
Administration	3	24.338***	.098	3	42.232***	.159
Administration X Treatment Group	6	1.077	.010	6	1.578	.014
Administration X Gender	3	.591	.003	3	.887	.004
Administration X Treatment Group X Gender	6	.377	.003	6	.440	.004
Error	672			672		
*p<.05 ** p < .01 *** p < .001						

Gender Scale Score Analysis

A 3 (Treatment: High-level, medium-level, No intervention) x 2 (Gender: male, Female) x 4 (GHSGT Spring Administration Scale Score Results: 2004, 2005, 2006, 2007) three-way mixed ANOVA was designed to test the effects of the treatments on gender and the average scale score of the school. As with Question one, scale score was used as a measure of science achievement. Descriptive statistics for each variable is shown below in Appendix F.

Mauchly's Test for Sphericity was found to be significant (W=.924, $p < .05$). However, the Greenhouse-Geisser Epsilon value is acceptable (.952). Therefore, sphericity can be maintained for the purposes of this study.

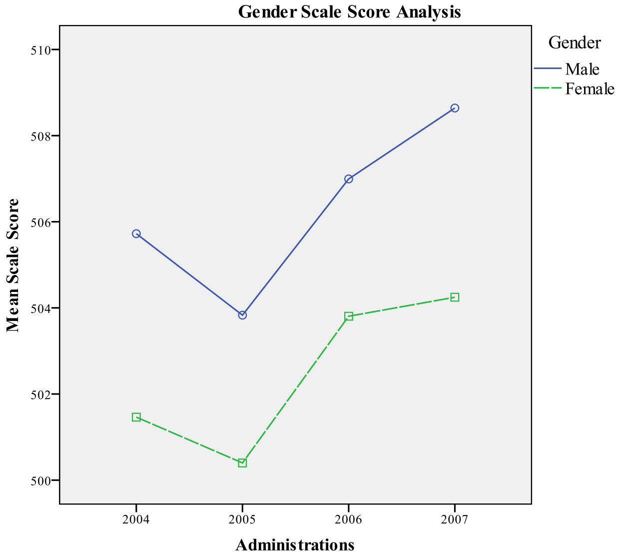
Tests of Within-Subjects Effects showed significant effects for each GHSGT Administration ($F_{2,672}$=24.338, $p<.001$). The study showed a strong power (1.000) and small effect size (Partial Eta Square=.159) indicating the effects are without Type I error. The significant effect was independent of Treatment Group ($F_{6,672}$=1.077), gender ($F_{3,672}$= .591) or interactions within the three ($F_{6,672}$= .377). Once again, 2004 was used as the baseline year due to no schools receiving treatment within that year. GHSGT Administrations in 2005, 2006, and 2007 all show a statistical difference to 2004 ($p<.05$) showing that all administrations saw an increase over the baseline year. A pairwise comparison shows significant differences in years 2005 ($p=.029$) and 2006 ($p<.001$). There was not a significant increase in 2007 ($p=.465$).

The Between-Subjects variable analysis showed no significant effect attributed to the Treatment Groups ($F_{2,224}$=1.194, $p=.963$) as shown in Table 6. Pairwise comparisons and Scheffe's post-hoc test also revealed no significant difference in scale scores between the subgroups. The analysis of gender did reveal a significant difference in mean scale score across all years between the performance of males and females ($F_{1,224}$=30.004, $p<.001$). The analysis showed a strong Power (1.000) and small effect size (Partial Eta Squared=.118) indicating the difference in scale score performance was strongly related. Pairwise comparisons and Scheffe's post-hoc tests confirmed the difference in scale score performance by gender. The analysis showed that regardless of treatment, males outperformed females during each administration as shown in Table 7 and plotted in Figure 6.

Table 7 – **Gender Mean Scale Score Gender Gap from 2004 and 2007**

Treatment Group	Subgroups	2004 Mean Scale Score	2005 Mean Scale Score	2006 Mean Scale Score	2007 Mean Scale Score
High-Level Intervention	Male	504.17	503.62	506.94	508.70
	Female	500.21	500.38	503.09	503.88
	Gap (Male-Female)	3.96	3.24	3.85	4.82
Medium-Level Intervention	Male	506.43	503.13	507.25	507.22
	Female	502.03	499.84	503.26	503.26
	Gap (Male-Female)	4.40	3.29	3.99	3.96
No Intervention	Male	506.57	504.74	506.79	510.00
	Female	502.14	500.98	505.07	505.60
	Gap (Male-Female)	4.43	3.76	1.72	4.40

Figure 6 – Gender Analyses Plot – Subgroup Scale Score Analysis



In 2004, high-level, medium-level and no-intervention schools exhibited an average achievement gap between male and female students of 3.96, 4.40, and 4.43 scale score points, respectively. By 2007, the gender gap for high-level, medium-level, and no-intervention schools were 4.82, 3.96, and 4.40, respectively. However, the treatment groups showed no statistically significant effects.

Gender Proficiency Analysis

A 3 (Treatment: High-level, medium-level, No intervention) x 2 (Gender: male, female) x 4 (GHSGT Spring Administration Percent Proficiency: 2004, 2005, 2006, 2007) was used to analyze percent proficiency. The descriptive statistics for each category are listed in Appendix G.

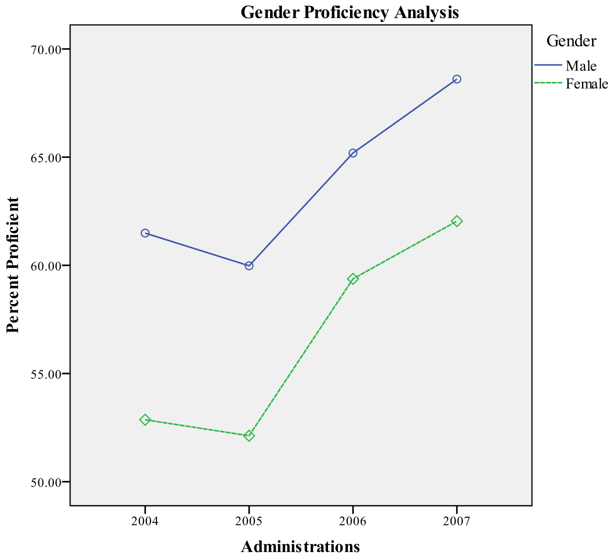
Mauchly's Test of Sphericity did show a significant effect (W=.924, $p<.05$). However, the Greenhouse-Geisser Epsilon value (.952) to allow for the assumption of sphericity to be maintained.

The Test for Within-Subjects revealed a significant difference in the percent of proficiency over administrations ($F_{3,672}$=42.232, $p<.001$) as shown in Table 6. The analysis displays strong power (1.000) and a small effect size (Partial Eta Square=.159) indicating the relationships are not by chance. Tests of within-subject contrasts show there are significant difference between the 2004 administration and those in 2006 ($p<.001$) and 2007 ($p<.001$). There is not a significant difference between the 2004 and 2005 administration (p=.252). A pairwise comparison shows significant differences from 2005 to 2006 ($p<.001$) and from 2006 to 2007 ($p<.05$). There is no significant difference from 2004 to 2005 (p=1.000). The analysis did not indicate that treatment group ($F_{6,672}$= 1.578) or gender ($F_{3,672}$= .887) had significant effect on proficiency. The tests of within-subject contrasts showed the treatment group to have a significant effect only between years 2004 and 2005 ($p<.05$), but no overall effect. This effect showed moderate power (.620) and a very small effect size (Partial Eta Squared=.029).

The analysis of Between-Subjects variables revealed a significant difference in gender ($F_{1,224}$= 25.938, $p<.001$), but not in Treatment Group ($F_{2,224}$= 2.441, p=.089) as shown in Table 6.

Observed power was strong with a small effect size with regard to gender (1.000 and .104, respectively). Plots of the differences in gender are located in Figure 7.

Figure 7 – Gender Analyses Plot – Male versus Female Proficiency Analysis



Pairwise comparisons confirmed the significant difference between the percent of proficient males and females. The gender gaps between male and female performance in percent proficiency increased in high-level intervention schools from 7.20 in 2004 to 8.67 in 2007 whereas the gender gaps closed in medium-level and no-intervention schools by 3.75 and 3.9 percentage points, respectively. This is an interesting effect given that 15 of the 16 mentors were female. The high-level intervention schools were below the state average in the gender gap in 2004 (8 percentage points) and above the state average in 2007 (6 percentage points). The opposite occurred for the medium-level and no-intervention schools with both beginning above the state average and finishing below the state average.

Ethnicity Analysis

The researcher performed a mixed analysis of variance where the GHSGT Administrations represented the within-subjects variable and the Treatment level and ethnicity (white; black; Hispanic; Asian) represented the between-subjects variables for both scale score and proficiency. For both analyses, significant findings were year to year performance and within the ethnicity subgroups. These results are summarized in Table 8.

Table 8 Ethnicity Analyses – Scale Score and Proficiency Mixed ANOVA Findings

	Scale Scores			Proficiency Scores		
	Df	F	Partial Eta Squared	df	F	Partial Eta Squared
Between Subjects						
Treatment Group	2	.044	.000	2	.014	.000
Ethnicity	3	77.857***	.404	3	88.679***	.434
Ethnicity X Treatment Group	6	.734	.013	6	1.138	.019
Error	345			347		
Within Subjects						
Administration	3	11.708***	.033	3	15.728***	.043
Administration X Treatment Group	6	1.718	.010	6	1.260	.007
Administration X Ethnicity	9	.836	.007	9	.935	.008
Administration X Treatment Group X Ethnicity	18	1.253	.021	18	1.550	.026
Error	1035			1041		
*p<.05 ** p < .01 *** p < .001						

Ethnicity Scale Score Analysis

A 3 (Treatment: High-level, medium-level, No intervention) x 4 (Ethnicity: White, Black, Hispanic, Asian) x 4 (GHSGT Spring Administration Scale Score Results: 2004, 2005, 2006, 2007) three-way mixed ANOVA was designed to test the effects of the interventions on the largest ethnic groups and the average scale score of the school. The descriptive statistics for the groups and categories are located in Appendix H.

Mauchly's Test of Sphericity showed a significant effect (W=.907, p<.05). However, Greenhouse-Geisser's Epsilon showed a sufficient value (.941) to allow for the assumption of sphericity.

The Tests of Within-Subjects Effects once again showed a significant effect on the mean scale scores over administrations ($F_{3,1035}$= 11.708, p<.001). The analysis showed sufficient power (1.000) and a very small effect size (Partial Eta Square=.033). According to the analysis, these effects were not the result of Treatment Groups ($F_{6,1035}$= 1.718, p=.118) or Subgroups ($F_{9,1035}$= .836, p=.577). The observed power and effect sizes were small for both of these interactions. Considering 2004 as the baseline year, the analysis showed significant contrasts between the baseline year and average scale score in 2006 (p<.001) and 2007 (p<.001). Both of these difference displayed strong observed power (.994 and .996, respectively) and small effect sizes (.055 and .058, respectively). The contrast also showed a significance interaction between scale score averages in 2004 and 2007 and the treatment groups (p<.05). This contrast yielded moderate power and a small effect size (.744 and .024, respectively). Lastly, there was a significant interaction for the same years between all three categories (p<.05). This contrast yielded a strong power and a small effect size (.823 and .041, respectively). A pairwise comparison revealed a significant differences between administrations in 2005 and 2006 (p<.05). There was no significant difference in any other years when compared to the previous year.

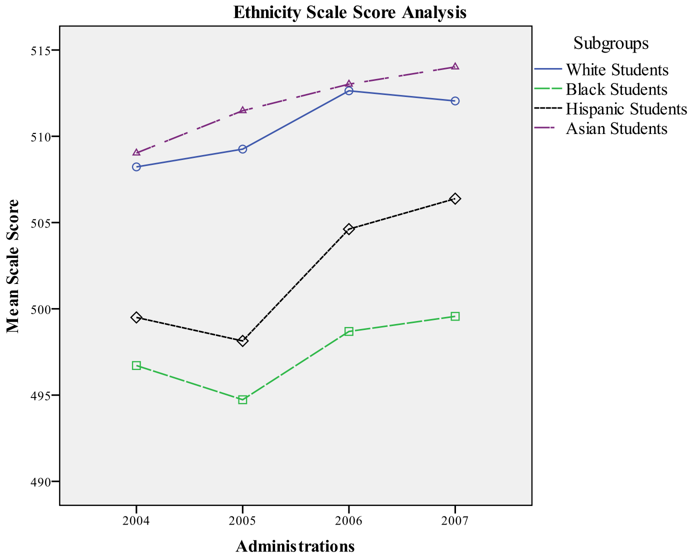
The Test of Between-Subjects Effects showed a significant effect in the ethnic subgroups ($F(_{3,345})$=77.857, p<.001) but no significant effect in Treatment Groups ($F(_{2,345})$=.044, p=.957).

Subgroups displayed a strong observed power and a small to moderate effect size (1.000 and .404, respectively) indicating the effects are not by chance. As can be seen in Figure 8, all ethnic subgroups experienced increases. However, white and Asian students performed significantly higher than black and Hispanic students as supported by the results shown in Table 9. It is important to point out the effect for ethnicity was a result of the average across all administrations.

Table 9 Ethnicity Scale Score Analyses – Overall Subgroup Means 2004 – 2007

Subgroups	Overall Scale Score Mean 2004 – 2008
White Students	510.541
Asian Students	511.888
Black Students	497.424
Hispanic Students	502.161

Figure 8 – Ethnicity Analyses Plot – Subgroup Scale Score Analysis



It is clear from this table that black and Hispanic students perform significantly different than their white and Asian counterparts. The achievement gap between white students and black and Hispanic students from 2004 to 2007 can be seen Table 10. While this data does show general trends in the scale score achievement gap, it is important to remember that there was no significant interaction between ethnicity and treatment or ethnicity and time.

Table 10 Ethnicity Analysis – Pre-Treatment – Post-treatment Scale Score Gap by Treatment Group

	High-Intervention	Medium-Intervention	No-Intervention
2004 White – Black Achievement Gap	14.6	7.26	12.68
2007 White – Black Achievement Gap	15.3	8.18	13.93
2004 White – Hispanic Achievement Gap	7.06	7.15	11.96
2007 White – Hispanic Achievement Gap	9.79	4.22	2.98

Regardless of treatment group, the number of scale score points that separate whites from blacks went up slightly whereas only the high-level intervention group went up with regard to the gap between white and Hispanic students. In the next section, an analysis of proficiency will determine if these scale score gaps translate into proficiency gaps.

Ethnicity Proficiency Analysis

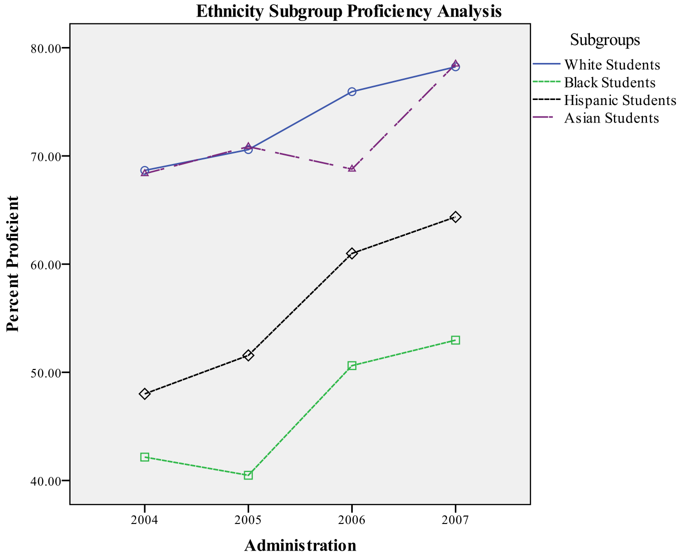
A 3 (Treatment: High-level, medium-level, No intervention) x 4 (Ethnicity: White, Black, Hispanic, Asian) x 4 (GHSGT Spring Administration Percent Proficiency Results: 2004, 2005, 2006, 2007) three-way mixed ANOVA was used to analyze the effects of the interventions on the largest ethnic groups and the average percent proficiency of the school. The descriptive statistics for the groups and categories are located in Appendix I.

Mauchly's Test of Sphericity showed a significant effect (W=.946, p<.05). However, Greenhouse-Geisser's Epsilon showed a sufficient value (.964) to allow for the assumption of sphericity.

The Tests of Within-Subject Effects show a significant effect in the percent proficiency from administration to administration ($F(_{3,1041})=15.728$, $p<.001$) as was seen in each of the other analysis. A summary of F-tests can be seen in Table 8. Significant differences over administrations is not due to Treatment Group ($F(_{6,1041})=1.26$, $p=.274$) or ethnic subgroup ($F(_{9,1041})=.935$, $p=.491$). Significant effect is supported with regard to administration by a strong observed power (1.000) but the analysis had a small effect size (.043). Again using 2004 as the baseline year, 2006 and 2007 show significance ($p<.001$) while there is no significant difference between 2004 and 2005 administrations. Significant differences in 2006 and 2007 are supported by a strong power (.942 and 1.000) and a small effect size (.035 and .087) for each administration. There are some interactions between 2004 and 2006 when analyzing the interactions within administration, treatment, and subgroups ($p<.05$). This is a significant interaction with strong power and a small effect size (.872 and .045, respectively). Pairwise comparisons show significant difference only between the 2005 and 2006 administration ($p<.05$).

Tests of Between-Subjects Effects show significant effects only with subgroups ($F(_{3,347})=88.679$, $p<.001$). This effect displays strong power and moderate effect size (1.000 and .434, respectively). There is no significant effect with regard to treatment groups or significant interactions between ethnic subgroups and treatment groups. As with scale score, white and Asian students exhibit significantly different proficiency rates than black and Hispanic students ($p<.001$) do. In addition, black students are significantly different than Hispanic students as well ($p<.001$). White and Asian students tend to have similar performance rates as shown in Figure 9.

Figure 9 – Ethnicity Analyses Plot – Subgroup Proficiency Analysis



Pairwise comparisons and Scheffe's post hoc tests confirm statistical differences in ethnic subgroups but no significant difference in treatment groups. The proficiency gap between white students and black or Hispanic students are listed in Table 11.

Table 11 Ethnicity Analysis – Pre-Treatment – Post-treatment Proficiency Gap by Treatment Group

	High-Intervention	Medium-Intervention	No-Intervention	State
2004 White – Black Proficiency Gap	32.73	18.77	28.00	31
2007 White – Black Proficiency Gap	27.80	22.18	25.82	25
Net Change between 2004 and 2007	4.93	-3.41	2.18	6
2004 White – Hispanic Proficiency Gap	15.67	20.47	25.83	34
2007 White – Hispanic Proficiency Gap	18.11	18.98	4.55	21
Net Change between 2004 and 2007	-2.44	1.49	21.88	13

High-intervention and no-intervention schools were able to lower their white-black achievement gap between 2004 and 2007. High-intervention schools saw a larger decrease in that time (4.93

percentage points) than no-intervention schools did. Medium-intervention schools actually saw an increase in the proficiency gap. Medium- and no-intervention schools saw their white-Hispanic proficiency gap decrease while the high-intervention schools saw a slight increase between 2004 and 2007. Just as with the scale score analysis, the high-intervention group saw improvement in 2007, but not at the same rate as the no-intervention group did.

Economically Disadvantaged Analysis

The researcher performed a mixed analysis of variance where the GHSGT Administrations represented the within-subjects variable and the Treatment level and subgroup (Economically Disadvantaged (ED) vs. non-ED) represented the between-subjects variables for both scale score and proficiency. In both analyses, there was a significant difference in performance from year to year, subgroup, and an interaction of subgroup and treatment group. These results are summarized in Table 12. Additionally, the researcher performed additional one-way ANOVAs to analyze the interaction of Treatment Groups with ED and non-ED students. For both scale score and proficiency, high-intervention ED students were found to be significantly different than medium- or no-intervention ED students only in the 2004 administration. This finding indicates high-intervention ED students performed statistically equivalent to the ED students in the other treatment groups from 2005 through 2007.

Table 12 ED Analyses – Scale Score and Proficiency Mixed ANOVA Findings

	Scale Scores			Proficiency Scores		
	Df	F	Partial Eta Squared	df	F	Partial Eta Squared
Between Subjects						
Treatment Group	2	22.963	.011	2	2.135	.020
Subgroup	1	4436.463***	.519	1	234.138***	.534
Subgroup X Treatment Group	2	4.821**	.045	2	3.915**	.037
Error	204			204		
Within Subjects						
Administration	3	30.037***	.128	3	43.417***	.175
Administration X Treatment Group	6	1.184	.011	6	1.812	.017
Administration X Subgroup	3	.243	.001	3	1.453	.007
Administration X Treatment Group X Subgroup	6	.210	.002	6	.432	.004
Error	612			612		
*p<.05 ** p < .01 *** p < .001						

Economically Disadvantaged – Non-Economically Disadvantaged Scale Score Analysis

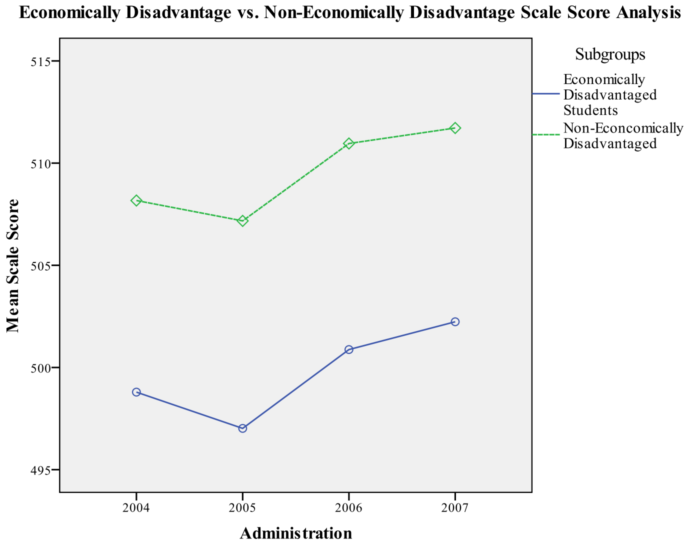
A 3 (Treatment: High-level, medium-level, No intervention) x 2 (Subgroup: Economically Disadvantaged (ED), non-ED) x 4 (GHSGT Spring Administration Scale Score Results: 2004, 2005, 2006, 2007), was designed to test the effects of the treatments on the average scale score of students identified as Economically Disadvantaged and non-Economically Disadvantaged. A student is identified as ED based on their Free and Reduced Lunch (FRL) status. Descriptive statistics for each variable is shown below in Appendix J.

Mauchly’s Test of Sphericity showed a significant effect (W=.935, p<.05). However, Greenhouse-Geisser’s Epsilon showed a sufficient value (.960) to allow for the assumption of sphericity.

Tests for Within-Subjects Effects show only significant effect in administrations over time ($F_{(3,612)}$=30.037, $p<.001$). This significant effect is supported by a strong power (1.000) and a small effect size (.128). The analysis shows no significant effect or interaction with the subgroups ($F_{(3,612)}$=.243, $p=.859$) or treatment groups ($F_{(6,612)}$=1.184, $p=.314$). All administrations display significant contrast against the baseline year of 2004 ($p<.05$). No other contrasts or interactions are significant. In a pairwise comparison, the analysis shows a significant difference between the administrations of 2005 and 2006 ($p<.05$).

Tests for Between-Subjects Effects show significant effects between subgroups ($F_{(1,204)}$=4436.463, $p<.001$) and the interaction between subgroups and treatment groups ($F_{(2,204)}$=4.821, $p<.05$). The observed power is strong for the subgroups effect (1.000) and moderate to strong for the interactions (.794). The effect size is moderate for the subgroup effect (.519), but small for the interaction effect (.045). The difference in performance between ED and non-ED can be easily seen in Figure 10.

Figure 10 – Economically Disadvantaged Analyses Plot – Subgroup Scale Score Analysis



Pairwise and Scheffe's post-hoc test confirms significant differences in the scale score performance of ED versus non-ED students. Pairwise comparisons confirm no significance between the treatment groups alone.

As can be seen in Table 12, the effect of the interaction between Treatment Group and Subgroups, while a small effect size (Partial Eta Square = .045) was significant. Means and standard deviations for these interactions are shown in Table 11.

Figure 11 – Economically Disadvantaged Analyses Plot – Subgroup x Treatment Group Scale Score Analysis

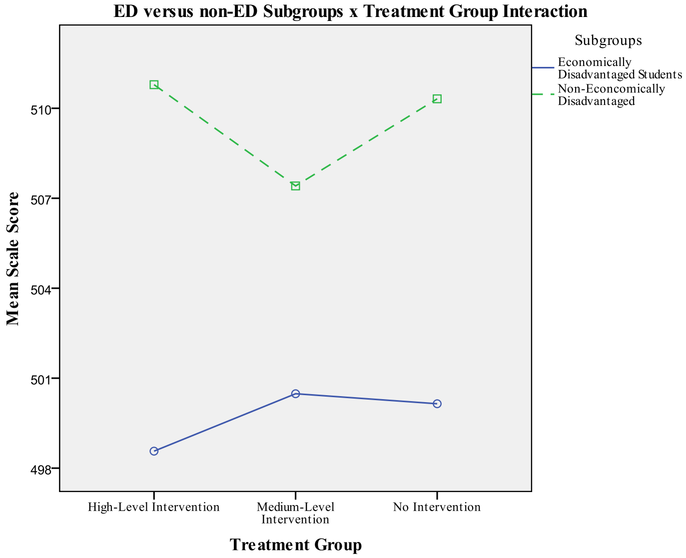


Table 13 – ED versus non-ED Scale Score Means and Standard Deviations by Treatment Group

Subgroup	No Intervention Mean (SD)	Medium Mean (SD)	High Mean (SD)
ED	500.147 (.669)	500.481 (.957)	498.566 (.648)
Non-ED	510.316 (.728)	507.411 (1.030)	510.788 (.728)

Since the effect size is so small, additional analyses were run to determine a more granular view of the effect. Independent T-Tests were used to analyze differences in subgroups by year. As can be seen in Table 14, all Treatment Groups showed significant differences in achievement for each administration between ED and non-ED students. Levene's Test for

Equality of Variance showed homogeneity for all samples except for Medium-level intervention schools in 2005 and 2006. For those years, there is still a significant difference in performance when homogeneity is not assumed. Homogeneity of variance assumes all populations have equal variance. An additional test, a one-way ANOVA would be a robust test even when homogeneity is not assumed as it does not require strenuous assumptions regarding the populations (Huck, 2000).

Table 14 – ED vs. non-ED Independent Samples T-Test Results by Administration

	Levene's Test for Equality of Variances		t-test for Equality of Means		
	F	Sig.	t	df	Sig. (2-tailed)
High Intervention Schools					
2004 GHSGT Scale Score	.040	.841	-10.698	86	<.001
2005 GHSGT Scale Score	.411	.523	-6.610	84	<.001
2006 GHSGT Scale Score	.008	.930	-11.254	84	<.001
2007 GHSGT Scale Score	1.565	.214	-10.306	85	<.001
Medium Intervention Schools					
2004 GHSGT Scale Score	.015	.903	-3.893	39	<.001
2005 GHSGT Scale Score*	4.455	.041	-3.114	28.715	.004
2006 GHSGT Scale Score*	5.501	.024	-4.219	31.978	<.001
2007 GHSGT Scale Score	1.358	.251	-4.149	40	<.001
No Intervention Schools					
2004 GHSGT Scale Score	2.174	.144	-6.706	81	<.001
2005 GHSGT Scale Score	1.716	.194	-7.317	84	<.001
2006 GHSGT Scale Score	.122	.728	-7.258	86	<.001
2007 GHSGT Scale Score	2.836	.096	-6.175	84	<.001

*Homogeneity of Variance not assumed

A one-way ANOVA analysis was used to compare scale score performance of ED students by Treatment Group and non-ED students by Treatment Group. Descriptive statistics can be seen in Table 15. The analysis of variance revealed the only significant difference in performance with regard to ED students was in 2004 (p<.05). A summary of the one-way ANOVA results can be seen in Table 16. Both Scheffe's and Tukey's post-hoc tests revealed a significant difference between high-intervention schools and medium- / no-intervention schools. While scale scores were still slightly lower in high-intervention schools, after the implementation of the SMP, there was no significant difference. No significant difference was found in the performance of non-ED students regardless of treatment or year.

Table 15 – ED and non-ED – Overall Subgroups Summary Scale Score Descriptive Statistics

		ED Analysis			Non- ED Analysis		
		N	Mean	SD	N	Mean	SD
2004 GHSGT Scale Score	High-Intervention	49	496.39	5.922	39	508.78	4.647
	Medium-Intervention	22	500.77	5.479	19	507.13	4.884
	No-Intervention	45	499.22	5.726	38	508.63	7.063
	Total	116	498.32	5.976	96	508.40	5.744
2005 GHSGT Scale Score	High-Intervention	48	495.95	10.191	38	508.58	6.617
	Medium-Intervention	22	497.79	4.668	20	504.53	8.582
	No-Intervention	46	497.47	6.390	40	508.76	7.912
	Total	116	496.90	7.941	98	507.83	7.690
2006 GHSGT Scale Score	High-Intervention	48	500.41	5.170	38	512.67	4.815
	Medium-Intervention	22	501.13	4.733	20	509.25	7.329
	No-Intervention	47	500.95	6.218	41	511.37	7.243
	Total	117	500.77	5.505	99	511.44	6.492
2007 GHSGT Scale Score	High-Intervention	48	501.51	5.724	39	513.20	4.618
	Medium-Intervention	22	502.22	4.844	20	509.12	5.922
	No-Intervention	46	502.89	6.584	40	513.38	9.105
	Total	116	502.19	5.918	99	512.45	7.134

Table 16 – ED & non-ED – One-way ANOVA Summary of Scale Score Results by Subgroup

		ED ANOVA Results			Non-ED ANOVA Results		
		df	F	Sig.	df	F	Sig.
2004 GHSGT Scale Score	Between Groups	2	5.274*	.006	2	.576	.564
	Within Groups	113			93		
	Total	115			95		
2005 GHSGT Scale Score	Between Groups	2	.598	.552	2	2.387	.097
	Within Groups	113			95		
	Total	115			97		
2006 GHSGT Scale Score	Between Groups	2	.173	.841	2	1.852	.162
	Within Groups	114			96		
	Total	116			98		
2007 GHSGT Scale Score	Between Groups	2	.634	.532	2	2.834	.064
	Within Groups	113			96		
	Total	115			98		

*p<.05

Economically Disadvantaged – Non-Economically Disadvantaged Proficiency Analysis

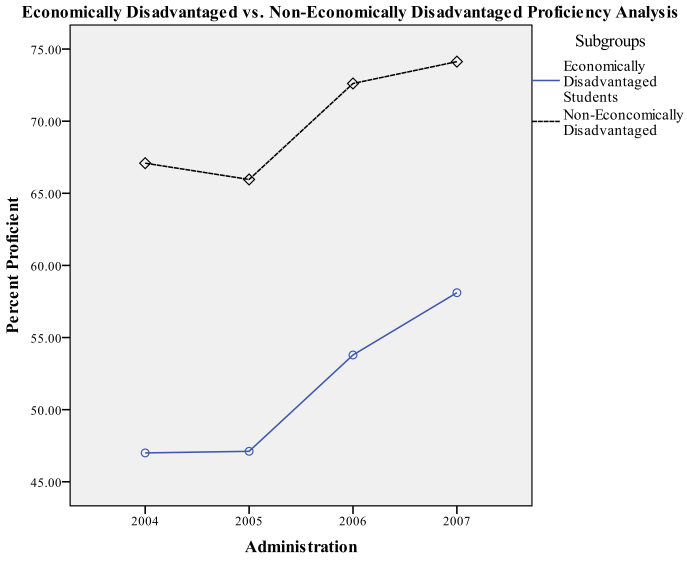
This proficiency analysis utilized a 3 (Treatment: High-level, medium-level, No intervention) x 2 (Subgroup: ED, non-ED) x 4 (GHSGT Spring Administration Percent Proficiency Results: 2004, 2005, 2006, 2007), was designed to test the effects of the treatments on the percentage of students proficient who are identified as Economically Disadvantaged (ED) and non-Economically Disadvantaged. Descriptive statistics for each variable is shown below in Appendix K.

Mauchly's Test of Sphericity showed a significant effect (W=.876, p<.05). However, Greenhouse-Geisser's Epsilon showed a sufficient value (.916) to allow for the assumption of sphericity.

The analysis of variance for the within-subjects effects of ED versus non-ED produced the same type of results as the previous analyses. There is significant effect in the administrations from year to year ($F_{(3,612)}$=43.417, p<.001) as seen in Table 12. Again, as with previous tests, the analysis shows a small effect size (.175) and strong power (1.000). There are no other significant effects with regard to interactions. There are significant contrasts when the 2006 ($p<.001$) and 2007 ($p<.001$) administrations are measured against the baseline year, 2004. There is a significant contrast by treatment group in 2005 ($p<.05$). Pairwise comparisons show a significant difference when the years 2006 ($p<.05$) and 2007 ($p<.05$) are compared with their preceding year.

The analysis of between-subjects effects shows significant effects among the subgroups ($F_{(1,204)}=234.138$, $p<.001$) and the interaction between treatment and subgroups ($F_{(2,204)}=3.915$, $p<.05$). The subgroup and interaction effects have sufficient power (1.000 and .701, respectively); however, the subgroups display a moderate effect size (.534) while the interaction displays a small effect size (.037). Pairwise comparisons confirm there are no significant differences with regard to treatment group. Pairwise and Scheffe's post-hoc confirm significant differences in ED and non-ED students. Those differences are also easily seen in Figure 12.

Figure 12 – Economically Disadvantaged Analyses Plot – Subgroup Proficiency Analysis



The differences in the proficiency gap for ED versus non-ED can be seen in Table 17. While this is important information to review for trend purposes, the significant difference was in overall effect as oppose to year to year.

Table 17 ED Analysis – Pre-Treatment – Post-treatment Proficiency Gap by Treatment Group

	High-Intervention	Medium-Intervention	No-Intervention	State
2004 ED – non-ED Proficiency Gap	26.29	13.95	20.02	26
2007 ED – non-ED Proficiency Gap	19.05	12.15	16.86	24
Net Change between 2004 and 2007	7.24	1.8	3.16	2

In all cases, the proficiency gap decreased. The proficiency gap in the high-intervention schools had the largest closer when compared to the other two treatment groups and the overall state average. Once again, there is a decrease in the rate of improvement in 2007 with regard to the high- and medium-level intervention schools. While the gap between ED and non-ED students closed more in high-intervention schools, this was not a significant difference.

Just as in the scale score analysis, additional tests should be administered to determine the effect of Treatment Group with ED students. The researcher used one-way ANOVA to analyze differences in ED students in the three Treatment Groups. Descriptive statistics can be seen in Table 18. The summary of results can be seen in Table 19.

Table 18 – ED and non-ED – Overall Subgroups Summary Proficiency Descriptive Statistics

		ED Analysis			Non- ED Analysis		
		N	Mean	SD	N	Mean	SD
2004 GHSGT Scale Score	High-Intervention	49	42.23	12.41	39	68.87	10.44
	Medium-Intervention	22	51.04	10.39	19	65.00	9.54
	No-Intervention	45	47.70	13.17	38	67.71	14.56
	Total	116	46.02	12.74	96	67.65	12.06
2005 GHSGT Scale Score	High-Intervention	48	47.03	11.17	38	69.07	10.09
	Medium-Intervention	22	46.59	10.30	20	60.95	14.70
	No-Intervention	46	48.16	10.21	40	68.54	14.29
	Total	116	47.39	10.56	98	67.20	13.17
2006 GHSGT Scale Score	High-Intervention	48	52.92	11.39	38	75.74	8.50
	Medium-Intervention	22	53.26	11.14	20	69.07	14.75
	No-Intervention	47	54.74	13.22	41	74.09	15.39
	Total	117	53.72	12.04	99	73.71	13.12
2007 GHSGT Scale Score	High-Intervention	49	56.54	11.39	39	75.89	7.80
	Medium-Intervention	22	57.27	7.73	20	69.58	12.41
	No-Intervention	46	60.37	10.14	40	77.69	11.47
	Total	117	58.18	10.37	99	75.34	10.72

The one-way ANOVA revealed that there was a significant difference between ED students in high-intervention schools and no-intervention schools the year before the program began in 2004 ($p<.05$). In the 2005 administrations and beyond, there were no significant differences. Again, while still having a lower proficiency rate, high-intervention ED students were performing statistically equivalent to their counterparts in the other treatment groups. Non-ED students were found to be statistically different in 2007; however, Scheffe's and Tukey's post-hoc test revealed the actual difference to be between medium-level and no-intervention schools.

Table 19 – ED & non-ED – One-way ANOVA Summary of Proficiency Results by Subgroup

		ED ANOVA Results			Non-ED ANOVA Results		
		df	F	Sig.	df	F	Sig.
2004 GHSGT Scale Score	Between Groups	2	4.527*	.013	2	.656	.521
	Within Groups	113			93		
	Total	115			95		
2005 GHSGT Scale Score	Between Groups	2	.212	.809	2	2.960	.057
	Within Groups	113			95		
	Total	115			97		
2006 GHSGT Scale Score	Between Groups	2	.288	.750	2	1.750	.179
	Within Groups	114			96		
	Total	116			98		
2007 GHSGT Scale Score	Between Groups	2	1.743	.180	2	4.151*	.019
	Within Groups	114			96		
	Total	116			98		

*p<.05

Students With Disabilities (SWD) Analysis

The researcher performed a mixed analysis of variance where the GHSGT Administrations represented the within-subjects variable and the Treatment level and subgroup (Students With Disabilities (SWD) vs. regular education students) represented the between-subjects variables for both scale score and proficiency. Significant findings for both scale score and proficiency were found in year to year performance, interaction between administration and subgroup, and subgroup differences. These results are summarized in Table 20.

Table 20 SWD Analyses – Scale Score and Proficiency Mixed ANOVA Findings

	Scale Scores			Proficiency Scores		
	Df	F	Partial Eta Squared	df	F	Partial Eta Squared
Between Subjects						
Treatment Group	2	.135	.003	2	.533	.005
Subgroup	1	418.089***	.652	1	874.743***	.813
Subgroup X Treatment Group	2	.087	.001	2	1.768	.017
Error	223			201		
Within Subjects						
Administration	3	8.1333***	.035	3	13.760***	.064
Administration X Treatment Group	6	.325	.003	6	2.023	.020
Administration X Subgroup	3	3.553***	.016	3	4.930**	.024
Administration X Treatment Group X Subgroup	6	.100	.001	6	1.723	.017
Error	669			603		
*p<.05 ** p < .01 *** p < .001						

Students with Disabilities (SWD) – Non-SWD Scale Score Analysis

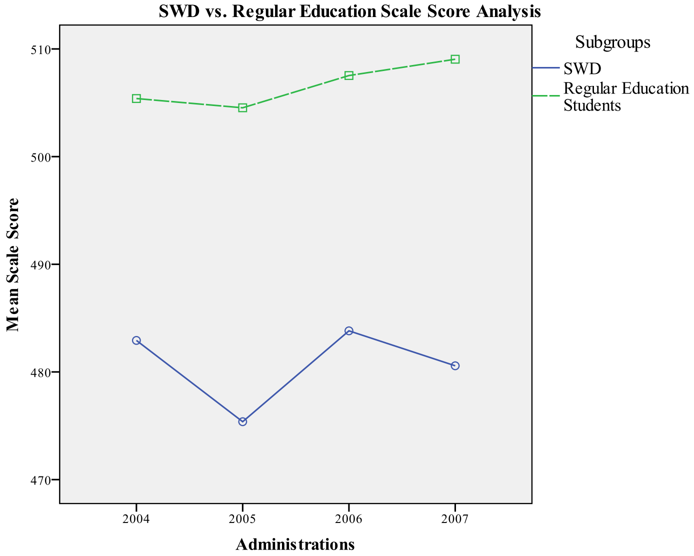
This scale score analysis utilized a 3 (Treatment: High-level, medium-level, No intervention) x 2 (Subgroup Average Scale Score: SWD, non-SWD) x 4 (GHSGT Spring Administration Scale Score Results: 2004, 2005, 2006, 2007), was designed to test the effects of the treatments on students' scale scores who are identified as Students with Disabilities (SWD) and non-SWD. Descriptive statistics for each variable is shown below in Appendix L.

Mauchly's Test of Sphericity showed a significant effect (W=.821, p<.05). However, Greenhouse-Geisser's Epsilon showed a sufficient value (.904) to allow for the assumption of sphericity.

This analysis of within-subjects effects displayed effects for administrations over time ($F_{(3,669)}$=8.133, p<.05) and subgroups ($F_{(3,669)}$=3.553, p<.001). No other effects were identified. Unlike previous analyses, only one administration, 2005, was statistically different (p<.05) than the baseline year of 2004. Observed power of this contrast was strong (.927), but it had a small effect size (.05). Pairwise comparisons revealed significant difference in the 2005 (p<.05) and 2006 (p<.05) when compared to the previous administration. The 2005 administration saw a significant drop in SWD. All three treatment groups saw significant drops in 2005. Two thousand and five was significantly different (p<.05 in all comparisons) to any other year in the study. There was a drop for all three groups in 2007.

The analysis for between-subjects effects yielded significant effects for only the subgroups ($F_{(1,223)}$=418.089, p<.001). The subgroup effect is supported by strong power (1.000) and a moderate to high effect size (.652). Pairwise comparisons show a significant difference in performance between SWD and non-SWD. This is also evident in the plot shown in Figure 13. Scheffe's post-hoc test also confirmed no other significant differences within the treatment groups.

Figure 13 – Students With Disabilities Analyses Plot – Subgroup Scale Score Analysis



As shown in Figure 13 and Table 21, performances of the two groups vary by administration. After experiencing a decrease in performance in the 2005 administration, the regular education students increased each year during the treatment. In addition, regular education students had little variation within the mean scale scores. SWD, on the other hand, saw a larger decrease in 2005, but increased in 2005 and 2006 before declining again in 2007.

Table 21 – SWD vs. Regular Ed. Students – Overall Subgroups Summary Descriptive Statistics

	Subgroups	N	Mean	Std. Deviation
2004 GHSGT Scale Scores	SWD Students	115	482.71	9.579
	Regular Education Students	116	505.31	5.996
2005 GHSGT Scale Scores	SWD Students	115	475.57	23.527
	Regular Education Students	116	504.68	6.642
2006 GHSGT Scale Scores	SWD Students	116	483.84	14.666
	Regular Education Students	117	507.60	6.260
2007 GHSGT Scale Scores	SWD Students	115	480.74	23.953
	Regular Education Students	116	509.32	6.031

Additional analysis through the use of independent sample T-Test shows that SWD consistently perform at a significantly lower level than regular education students ($p<.05$). A summary of the T-Test findings are shown in Table 22. Each administration was significantly different.

Homogeneity of variance was not assumed in any case. The mean difference shown in Table 20 again represents that the greater difference between the two groups could be seen in 2005 and 2007.

Table 22 – SWD vs. Regular Ed. Students – Summary of T-Test Findings by Administration

	Levene's Test for Equality of Variances		t-test for Equality of Means			
	F	Sig.	t	df	Sig. (2-tailed)	Mean Difference
2004 GHSGT Scale Scores	21.936	<.001*	-21.474	191.178	<.001	-22.602
2005 GHSGT Scale Scores	18.502	<.001*	-12.772	131.911	<.001	-29.108
2006 GHSGT Scale Scores	10.085	.002*	-16.057	155.276	<.001	-23.758
2007 GHSGT Scale Scores	13.814	<.001*	-12.412	128.279	<.001	-28.582

*Homogeneity of variance not assumed

Students with Disabilities (SWD) – Non-SWD Proficiency Analysis

This proficiency analysis utilized a 3 (Treatment: High-level, medium-level, No intervention) x 2 (Subgroup Average Percent Proficiency: SWD, non-SWD) x 4 (GHSGT Spring Administration Scale Score Results: 2004, 2005, 2006, 2007), was designed to test the effects of the treatments on students' percent proficiency who are identified as Students with Disabilities (SWD) and non-SWD. Descriptive statistics for each variable is shown below in Appendix M.

Mauchly's Test of Sphericity showed a significant effect (W=.841, p<.05). However, Greenhouse-Geisser's Epsilon showed a sufficient value (.904) to allow for the assumption of sphericity.

The analysis of the within-subjects effects showed significant effects in two areas: administrations ($F(_{3,603})$=13.760, p<.001) and subgroups ($F(_{3,603})$=4.930, p<.05). A summary of F-test results are shown in Table 20. Both effects display strong power (1.000 and .888, respectively), however, both also show small effect sizes (.064 and .024, respectively). When analyzed against the baseline year, 2004, all other administrations showed significant differences ($p<.05$). In addition, the contrasts showed significance in the interactions between administrations and treatment groups during the 2007 administration ($p<.05$). There were also significant differences in the interactions of administrations and subgroups during the 2006 ($p<.05$) and 2007 ($p<.05$) administrations when compared to the baseline year. Finally, there was a significant interaction between all three variables when 2007 ($p<.05$) was contrasted with the baseline year. In general, the power levels for all of the contrasts were moderate to strong as seen in the Table 23. The effect sizes were all small for the contrasts. A pairwise comparison showed a significant comparison between 2005 and 2006 ($p<.001$).

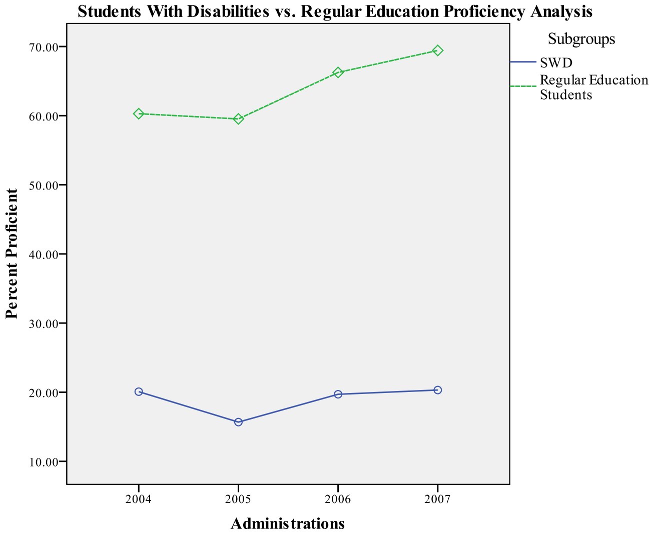
Table 23 SWD Analysis – Tests of Within-Subjects Contrasts for SWD Proficiency Analysis

Source	Administrations	df	F	Sig.	Partial Eta Squared	Observed Power[a]
Administrations	Level 2 vs. Level 1	1	5.832	.017	.028	.671
	Level 3 vs. Level 1	1	3.916	.049	.019	.504
	Level 4 vs. Level 1	1	15.181	<.001	.070	.972
Administrations * TreatmentGroup	Level 2 vs. Level 1	2	2.688	.070	.026	.528
	Level 3 vs. Level 1	2	2.110	.124	.021	.430
	Level 4 vs. Level 1	2	5.829	.003	.055	.868
Administrations * Subgroups	Level 2 vs. Level 1	1	2.914	.089	.014	.397
	Level 3 vs. Level 1	1	5.037	.026	.024	.608
	Level 4 vs. Level 1	1	13.695	<.001	.064	.958
Administrations * TreatmentGroup * Subgroups	Level 2 vs. Level 1	2	.311	.733	.003	.099
	Level 3 vs. Level 1	2	.991	.373	.010	.221
	Level 4 vs. Level 1	2	4.366	.014	.042	.751

a. Computed using alpha = .05

The analysis of the between-subjects effects showed only significant effect for the proficiency of SWD versus regular education students ($F_{(1,201)}=874.743$, $p<.001$). The power and effect size were strong for this effect (1.000 and .813, respectively). Pairwise comparisons confirmed a significant difference in the performance of SWD and regular education students (p<.05). The plot of the proficiency rates for each group can be seen in Figure 14.

Figure 14 – Students With Disabilities Analyses Plot – Subgroup Proficiency Analysis



Students in the high-intervention schools exhibited lower proficiency rates in 2004 than either of the other two groups. While all groups decreased in 2005, the high-intervention students did not see as large a decrease as the other two groups. For the purposes of this study, Table 24 below shows the changes in the proficiency gap between 2004 and 2007 as produced in previous analysis. 2005 is being added for the purpose of viewing the changes from the inception of the program. As with the SWD scale score analysis, schools saw a decrease in proficiency in 2005. Unlike the scale score analysis, however, proficiency rate continued to increase in 2007.

Table 24 SWD Analysis – Pre-Treatment – Post-treatment Proficiency Gap by Treatment Group

Subgroup	2004 Mean (SD)	2005 Mean (SD)	2006 Mean (SD)	2007 Mean (SD)
Students With Disabilities	19.4111 (15.65733)	16.0081 (13.54748)	20.6721 (17.90123)	22.2364 (18.24421)
Regular Education Students	60.0322 (13.35598)	59.9705 (12.34186)	66.5431 (12.91237)	69.8265 (10.87936)

As can be seen in Table 24, only the high-level intervention schools saw a decrease in the proficiency gap. As shown in the descriptive statistics and the plot in Figure 22, high-intervention started with a lower rate in 2004 and finished with a higher rate in 2007. This analysis showed the lone exception to a decrease in improvement rate in 2007.

Summary

In the analysis of these two research questions, significant differences in the within-subjects variable were found in administrations of the GHSGT science portion in every case. The between-subjects variables found to be statistically significant were differences among subgroup performance. Specifically, there were performance differences in the scale score analysis for male versus female; white/Asian versus black/Hispanic; ED versus non-ED; and SWD versus regular education students.

There was a significant effect between treatment group and subgroups with regard to ED for both achievement and proficiency rate. Additional tests showed ED in high-intervention schools only showed significant differences in performance to the other two groups in 2004. ED students in high-intervention schools have constant improvement between 2004 and 2007. While the ED population in high-intervention schools has consistently performed at a lower level when compared to the other two treatment groups, they have not been significant differences. ED students consistently performed significantly lower than non-ED students. While the proficiency rate among SWD at high-intervention schools did surpass the other treatment groups, it was not significant.

Chapter V. Discussion and Conclusions

Overview

In 2005, the Georgia Department of Education (GaDOE) began the implementation of a new set of high school science standards, Georgia Performance Standards (GPS). These standards were based on conceptual science and its practices. Given the poor performance on the GHSGT science portion throughout the years, the SMP was designed to focus schools on improvement and on quality practices of inquiry and instruction. This study was conducted to review the effectiveness of a state-level initiative known as the Science Mentor Program (SMP) designed to support teachers and improve student achievement in science. The study used two different metrics to determine effectiveness: 1) performance on the science portion of the Georgia High School Graduation Test (GHSGT) as measured by each school's average scale score and 2) proficiency on the science portion of the GHSGT as measured by the percent of students in each school meeting or exceeding the standard. These two metrics were used to analyze differences in year to year performance and to compare these performances to three groups of schools receiving different levels of intervention by the SMP. The Science Mentors were trained to implement the new standards and to implement quality science practices in high schools with the intent of improving the achievement and proficiency on the GHSGT.

A key factor in improving science achievement was to focus on improved performance in particular subgroups who have traditionally exhibited lower achievement on the science portion of the GHSGT. Of particular interest in both the SMP and this study, were Economically

Disadvantaged (ED) and Students With Disabilities (SWD) science achievement and proficiency. With the emphasis on science as inquiry with its requirement for more advanced thinking skills, it was hypothesized that, these two subgroups and their teachers might need additional support to improve performance.

Some studies have indicated that when science is taught using inquiry-based methods, students gain achievement, regardless of race, gender, or socio-economic situation. Lee and colleagues (2006) found that all students benefited from inquiry, but it was particularly important for ED and SWD. SWD and ED students have traditionally performed less than their peers on state, national, and international assessments (GaDOE, 2009; NCES, 2009; Bybee, 2009). Mastropieri & Scruggs (2006) state that students from underprivileged and non-mainstream backgrounds perform lower due to lack of science experience at earlier grade levels and often a lack of social context for the science content. One way to close the achievement gaps is to allow students to construct their own knowledge through inquiry (Basu & Barton, 2007; Driver, 1989). One such study found that when typical classroom instruction is compared to use of inquiry in instruction, there is no significant difference in performance in gender, ethnicity, or ED (Wilson, Taylor, Kowalski, & Carlson, 2010). That is to say that there was no significant achievement gap between subgroups when inquiry is used as opposed to the traditional classroom instruction where there was a significant achievement gap between subgroups of students. Lee and colleagues (2006) found that while student achievement increased with inquiry focused instruction, students from non-mainstreamed or less privileged backgrounds showed much higher gains than their mainstreamed, more privileged counterparts. Inquiry is perhaps a great equalizer. Students tend to be more engaged, interested and perform at higher levels when presented with inquiry activities that require them to use their minds to construct knowledge rather than always engaging in more traditional forms of instruction such as lecture and worksheets.

It was the contention of the GaDOE staff that for the SMP to have an impact, instruction and achievement for all students needed to improve. The GaDOE focused on strategies that would improve the performance of subgroups and thereby improve statewide performance. Most of the rural, and some urban areas, had little resources or expertise in the area of inquiry or instructional strategies that impacted students who may have previously been denied quality science education experience or access. The individuals hired had displayed quality instructional practices and student achievement outcomes while they were in the classroom.

Overall, the average scale score state wide on the science portion of the GHSGT was 506 (500 is needed to meet standards) and the percent of students meeting or exceeding the standard were at 68% in 2004. In this study, the researcher focused on the set of schools identified in the first year of the SMP (N = 48) that received high-intervention for a minimum of three consecutive years. The scale score and proficiency rates of these schools were analyzed over four years of data (2004 – 2007) for both progress and against a set of schools receiving medium-level intervention (N = 22) and schools that received no-intervention (N = 45). In 2004, the scale scores for the high-intervention schools were nine scale score points behind the state average, yet only four, behind the medium- and no-intervention schools. In addition, they were 14 percentage points below the state average in proficiency rate. The GaDOE found in 2004, that schools within the metro Atlanta area tended to have a higher proficiency rate than schools outside the metro area (Pruitt, 2005). For this reason, the SMP focused more on schools outside the metro Atlanta area. Four Atlanta schools were selected to receive service based on extreme need and to evaluate impact within an urban setting. In general, SMP schools saw significant increases between 2004 and 2007 in both scale score and proficiency rates. All schools in this study regardless of intervention level saw an increase in both measures.

Summary of Key Findings

Research Question 1

Is there a significant difference in science achievement and proficiency for schools supported by the SMP each year in performance, between SMP and comparable schools, and between SMP schools receiving medium- versus high-level intervention on the science portion of the GHSGT from 2004-2007 during the period of intervention?

As a result of these analyses, the null hypothesis for year to year improvement can be rejected. There was a significant increase in achievement and proficiency rate from year to year within the SMP schools. However, the null hypothesis is confirmed with regard to comparisons of treatment groups. The analysis did not show significant improvement in achievement or proficiency rate between high-, medium-, or no-intervention schools. All three subgroups showed an increase in achievement and proficiency between 2004 and 2007 with a large increase in 2005, but no group showed significance to another.

Research Question 2

Is there significant improvement in the science achievement and proficiency on the science portion of the GHSGT for subgroups (male, female; Economically Disadvantaged (ED), non-Economically Disadvantaged; students with disabilities (SWD), non-SWD; White, Black, Hispanic, Asian) within schools receiving high-level intervention by the SMP and between SMP and comparable schools from 2004-2007?

As a result of these analyses, the null hypothesis for year to year improvement can be rejected. There was a significant increase in achievement and proficiency rate from year to year within subgroups of the SMP schools. In addition, there were also significant differences in achievement and proficiency with regard to the difference between male and female, ethnic groups, Economically Disadvantaged (ED), and Students With Disabilities (SWD). Specifically males outperformed females, whites and Asian students outperformed blacks and Hispanics, non-ED outperformed ED, and regular education students outperformed SWD.

The null hypothesis can be rejected for the effect the SMP had on ED. There was a significant difference in scale score and proficiency in ED in high-intervention schools when compared to schools receiving no-intervention. For both the scale score and proficiency rate analyses, a significant interaction between treatment groups and subgroups were found. Upon further testing, the significant difference was determined to be a significant difference in performance between ED and non-ED in 2004, the year prior to the implementation of the SMP. After implementation, high-intervention schools' ED population was statistically equivalent to medium- and no-intervention schools. This indicates there was a positive effect by the SMP on the performance of ED students. The ED versus non-ED scale score gap did decrease slightly over time; however, the standard deviation for ED decreased as well. There was also a substantial decrease in the proficiency gap in high-intervention schools. While the overall mean scale scores were lower than mediuim- and no-intervention schools, more students were scoring at the proficient level on the GHSGT.

The null hypothesis is confirmed with regard to other subgroup performances. As with Research Question 1, while there were general upward trends, the treatment groups revealed no significant differences in achievement or proficiency rate.

While the majority of analyses in this study resulted in no significant difference, there were areas of improvement. Indicators of such are

- All students in the high-intervention schools showed an increase in proficiency rate during the first year of implementation of the Georgia Performance Standards (GPS) in 2005 as opposed to a drop by no-intervention schools.
- All students in high-level intervention schools started 4.05 percentage points behind the no-intervention schools in 2004 and slightly closed that gap through 2007 (3.99 percentage points)
- Gender analysis showed high-intervention schools did not have as dramatic a drop in achievement in 2005 as no-intervention schools. Proficiency rates actually increased in high-intervention schools for 2005.
- Ethnicity and ED analyses showed a greater increase in achievement and proficiency between 2005 and 2006 in high-intervention schools than no-intervention schools.

Possible Explanations of Findings

Year to Year Improvements

For both research questions, all groups showed an increase from year to year. In almost all cases, there was an initial drop in both achievement and proficiency rate. In discussions with assessment and curriculum staff from the GaDOE, the prevailing thought is the science portion of the GHSGT experienced an "implementation dip" in 2005 brought on by the implementation of new standards and a new assessment. With new standards being implemented in the 2004-2005 school year, students were exposed to a new requirement in science as described in Chapter 1. Items were developed using both the Characteristics of Science and content standards requiring less memorization and more application. In short, the old standards, Quality Core Curriculum (QCC), employed standards that lent themselves to assessment items of discrete, low-level knowledge and fact-based responses. The Georgia Performance Standards (GPS), while more cognitively complex lent themselves to more conceptual and application-based assessments. Students under the QCC were required to remember small facts from as long as three years prior to the assessment whereas the GPS operated on the enduring understandings and big ideas in science which allowed for more use of knowledge.

Differences in Subgroup Performance

The study supported findings that have been historically reported with regard to disparity of subgroup performance. Males, white, non-ED, and regular education students were all found to have significantly better performance in both achievement and proficiency rate. These differences are most likely due to documented reasons such as access to quality instruction, resources and prior experiences in science. Equity in science instruction and assessment has been a concern for a long time. Underserved and underperforming students tend to be students of color, disabilities, or in poverty. Opportunity to Learn (OTL) is a key factor when discussing underserved or minority students. OTL is both a legal and professional issue with regard to the valid use of test results (Penfield, & Lee, 2010). ED students and SWD tend to have their opportunity to learn limited in several ways including teacher quality and curricular resources (Darling-Hammond, 2004). The intent of the SMP was to improve science achievement across Georgia to help more students meet proficiency on the science portion of the GHSGT. One key strategy is to close gaps between these subgroups using best practices and opening access to all students. More years of study will be needed to determine if this approach will have any significant effect on the closing of the achievement gaps.

Economically Disadvantaged Subgroup Performance

As has been stated earlier in the study, the GPS and SMP were both designed to engage students in meaningful and conceptual science. It is beneficial for all students, but ED have been traditionally left out of inquiry experiences due in large part to differences in social contexts (Lee, & Luykx, 2005; Lee, et.al., 2006). According to Lee and colleagues, traditional instruction makes assumptions about language and context that do not motivate or interest ED students. The Driver and Associates theory (1989) focuses on students having prior knowledge and adjusting that knowledge in the face of new data. However, ED students may not connect that prior knowledge if they are not motivated to engage. It is beneficial for ED to engage in activities that will enhance their inductive and deductive thinking skills and help them formulate their own ideas about the world around them. For students from low-income families and minorities, learning science in school is not as simple as the content. Learning school science is as much about understanding the language and context in which science is learned as it is the content of science itself (Barton, & Tan, 2009). Warren, Ballenger, Ogonowski, Rosebery, and Hudicourt-Barnes (2001) found that students of low-SES and minorities use their own experiences to communicate their context for learning science. Warren and his colleagues argue that students from diverse backgrounds should be acknowledged for the science connections they make to everyday life. Warren states,

> We are arguing for the need to analyze carefully on one hand the ways of talking and knowing that comprise everyday life within linguistic, racial, and ethnic minority communities, and on the other, the ways of talking and knowing characteristic of scientific disciplines (recognizing that even here there are important differences, say, between modes of explanation in physics and evolutionary biology).

In essence, alternative conceptions students have regarding the natural world are typically the result of their understanding of their own experiences. Teachers should also realize that those conceptions may be a different way of knowing. Engaging students, regardless of subgroup, where they are is paramount. Teachers allowing students to use previous knowledge engages and interests them in the science and can cause the student to become motivated to learn and express him or herself in a more scientific arena. Inquiry allows students more interaction with peers and teachers which enhance the learning experience.

Students of all ages come with misconceptions in science. These misconceptions can be the result of life experience or a misinformed explanation. A key strategy to remove these misconceptions is through the use of inquiry to cause them to question their own understandings ((Llewellyn, 2002). Inquiry engages students who have low experiences with the vocabulary of science in communicating science through their experiences and in the process, building their own scientific vocabulary (Barton, & Tan, 2009).

High school science teachers have traditionally not been trained in the art of differentiating a classroom. The GaDOE set inquiry and differentiation as a core function for the SMP. They were then trained in differentiation of instruction and the GaDOE's own model of Response to Intervention (RTI). The SISs spent time together and in a statewide collaborative known as the MSP-RESA Collaborative which was funded by a National Science Foundation Grant that brought K-12 and Higher Education together to improve science instruction. The SIS worked with other GaDOE staff, Regional Education Service Agency (RESA) science staff, and science supervisors to develop and implement methods of effective inquiry instruction and how to use it with the diverse population of learners represented throughout the state. ED in schools receiving high-intervention from the SMP achieved significantly lower than ED students in schools receiving middle- or no-intervention SMP during the first year of implementation. After the first year, even though the overall score remained lower, there was no statistically significant difference meaning that ED students from high-intervention schools were performing at an equivalent level to ED students in medium- or no-intervention schools. One reason for this could be the high level of support and coaching provided to the teachers and this connection should be investigated further. While it could be argued that the SMP did not have significant effect from 2004 through 2007 when compared to schools who received no service, it can be equally argued that this intervention engaged and impacted ED students. More ED were engaged in science, more ED students were given access to quality, rigorous inquiry instruction, and as a result more ED students are eligible to graduate high school.

Implications of Findings

This study was to develop an understanding of the impact the SMP has had on schools that received services for at least three years. There are four implications that can be derived from this study. First, with regard to SMP schools improvement, this study showed that all schools improved in both science achievement as defined by the school's average scale score on the science portion of the GHSGT and on the proficiency rate as defined by the percent of first time test takers meeting or exceeding the standard set for the science portion of the GHSGT. Each of the other two groups used in this study as well as the overall state average saw an increase over the four years included in the study. The researcher believes this increase is due in part to statewide support of the SMP, but also to a more focused set of science standards with a greater alignment to the GHSGT. After consulting the Center on Education Policy (CEP) and other literature, data on science achievement is limited at the state level. Many states did not have a high school assessment in science prior to the requirement under *No Child Left Behind* which required states to assess science once in grades 3-5, 6-8, and 9-12. More data will need to be collected as states continue their high school science assessment.

Second, the SMP had a significant effect on the scale score and proficiency rates of Economically Disadvantaged (ED) students. While the ED in the high-intervention group still performed lower overall when compared to the other two treatment groups, they were only significantly different in 2004 prior to the treatment. ED students still showed steady increases throughout the treatment. This finding should provide the GaDOE with some evidence of impact as well as provide a foundation to build additional supports for other subgroups.

The third implication is the SMP does provide a new model of statewide support designed to improve science instruction and achievement. This study, while showing significant effects for only a small proportion of students, does take the first step toward evaluating programs that intend to improve science education through a combination of state-level policy and grass-roots classroom impact. Science instruction and achievement have always been the focus of science educators. One possible solution in the U. S. is to facilitate state departments of education to collaborate with teacher preparation programs and local school to find a common ground to build science reform as it relates to teachers' quality and preparedness to implement a rigorous, inquiry-based curriculum. A program like the SMP that permeates the field with individuals who can act as support to teachers in need and act as conduits of communication to state departments of education and teacher preparation programs may be one answer to finding a common vision for the reform. Teachers receiving real-time, job-embedded professional development on inquiry and differentiation of instruction provide a much stronger context for teachers and prepare struggling teachers to manage on their own (Melville, & Bartley, 2010; (Peressini, Borko, Romagnano, Knuth, & Willis, 2004). In addition, the perceived success of the SMP has prompted the GaDOE to initiate a similar program in mathematics for the implementation of the new high school mathematics standards. However, due to budget constraints, implementation will prove more difficult as the number of positions in the Math Mentor Program is significantly fewer, but this same type of study should be conducted as soon as enough quality data exists.

The fourth implication is a definite need for additional study. Because the 2005-2006 school year included the implementation of several factors that could have affected increases in science achievement and proficiency between the 2005 and 2006 spring administrations, more data and study is needed to evaluate the full significance of the SMP over time. Each analysis showed an increase in overall score from year to year within all subgroups and closing of gaps within those subgroups and with the state averages. This could have been a function of the new GPS being implemented, but schools receiving high-intervention from the SMP did not see as large a drop in most subgroup performance within the first year of implementation (2005) as did the schools receiving no intervention. In general, high-intervention schools closed on the state average more than no-intervention schools. So, there are trends that indicate the SMP has impacted achievement; however, more study is needed. The GHSGT proficiency score was revised in 2008, another study of this kind would be beneficial to review its effectiveness on the new assessment. It would also appear, based on the results of this analysis that use of the SMP shows substantial return in the first year, but does not necessarily show a continued improvement over an extended period of time. One reason for this could be the influx of new ideas and willingness to try new things in the first year with a GaDOE employee in the room. There was certainly some trepidation on the part of teachers when they found out the GaDOE was sending someone to their school to support them. Many teachers felt it was to monitor and perhaps remove them. They tended to do things as suggested by the SISs, even though some did it begrudgingly. Over time, the relationship developed between the mentor (SIS) and the classroom teachers. As that relationship and comfort level with inquiry developed, the teachers began to exert more initiative. Perhaps the increase of the first year is the result of either wanting to impress staff or the fear not too. A potential recommendation going forward could be to provide only high-level support for one year and medium-level in the next. However, if the schools have a significant number of ED, the recommendation would be to continue high-level support.

While the GaDOE has valued this program for its contributions to overall science improvement in Georgia, the SMP stands in jeopardy in the current economic environment. Two-thousand, nine marked the first time in the history of the science GHSGT surpassed another content area in proficiency. Since the transition to a fully new GHSGT in 2008 and effective implementation of the GPS in science, statewide average proficiency rate in 2010 has reached 90% for all first-time test takers equaling English-Language Arts (ELA) in proficiency rate on the GHSGT. In addition, 57% of first-time test takers scored at the Advanced Proficiency or Honors level also equaling ELA. In 2010, the results of the science GHSGT were equal to the highest proficiency rates for all content areas. In 2003, only 41% of African-American students scored at a proficient level on the science portion of the GHSGT. In 2010, that proficiency rate has gone up to 79% proficient. This accounts for 19,176 more African-American students on track for graduation than in 2003. This success, coupled with a continued economic crisis has put the program at risk.

Limitations of the Study

It is difficult to measure the full impact of such a program due to the diversity of services it brings. Many of the Science Implementation Specialists (SIS) also conducted summer workshops on inquiry for all the systems in their area as well as trained teachers from schools and districts they were not assigned to in GPS. A major limitation is SIS were never meant to be limited to just their high-level or medium-level intervention schools. This provides a key limitation in that most of the other schools in that area were probably impacted by a workshop or training the SIS provided.

The school-level data serving as the unit of measure is certainly a limitation of the study. Because of the small sample sizes and fairly homologous schools, there was not a great deal of variability which most likely led to small and moderate effect sizes. Because this study focused on the results of the GHSGT, as this was the metric used to procure the funding for the program, it does limit the ability to review impact of the program. Due to the large nature of a fledgling program, the action plans developed for each school were as diverse as the schools themselves. As a first study of this program, the researcher chose to focus on the overall indicator of the schools' success. Future studies should also include analysis on the End-Of-Course Tests (EOCT). This researcher chose not to use this as the focus was on the actual end of high school assessment, but EOCT may show more direct and specific impact at the school level.

A limitation to this study is the small sample size of Hispanic and Asian students in the schools addressed in this study. While white and black students were fairly stable in terms of their percent of the schools population, Hispanic and Asian students were more subject to larger changes as a result of the population changing by even one to two students. Native American and multicultural had to be eliminated from the study due to such a small number of students and schools containing enough of those students to have an official subgroup.

Another area for future study is the Economically Disadvantaged. ED displayed statistical differences within the treatment groups. However, while there was a significant interaction between the subgroups and treatment groups, the effect size was so small (Partial Eta Squared <0.1), additional study would be required to boost the effect size to an acceptable level to determine if there was a powerful statistical significance with regard to an ED and treatment group interaction. In addition, more longitudinal data is needed to determine full impact on ED students.

The SMP is also impacted by the attrition of administrators and teachers in the schools they serve. Many times since the inception of the program, two to three teachers would leave a given school. While this does not sound like many, it is significant in a small rural school district that only has 4 science positions. In addition, a key requirement to success was a supportive administrator who understood their role as well as the role of the SIS. In discussions with the SIS, they resoundingly stated that their most successful schools were the ones in which the administrators understood they should hold the teachers accountable for the school science action plan and not expect the SIS to evaluate or act as an administrator advocate while working with the teachers. So, each time an administrator changed, the process of developing relationships began again.

The number of years of GHSGT results was also a limitation. As stated earlier, the GaDOE reset the minimum scores (cutscores) needed to meet the standard on the GHSGT. In doing this, it interrupted the ability on the part of a researcher to reliably compare tests prior to 2008 and after. The new scale score minimum was set at 200. The state has also seen a large upswing in the proficiency rate since 2008. This study could bear repeating in 2011 if the program survives the current economic climate.

Suggestions for Future Study

Based on the increases seen in recent years in science achievement and proficiency on the GHSGT in science, this program should be studied further to determine its impact. Due to the nature of the work of the SMP, a comparison of the action plans and schools' areas of focus would be helpful in the evaluation of the program. As discussed in earlier chapters, the SMP works with the high-intervention school to determine areas to focus that are as broad as Biology and Physical Science courses and as specific as increasing the percentage of time spent using inquiry. A study could involve both qualitative and quantitative components that focus on the increases in teacher efficacy with regard to their pedagogical content knowledge and their comfort using inquiry instructional or assessment tasks and analyze the effects on the specific End of Course Test the action plan set as the area of focus. The Biology and Physical Science End of Course Tests would be a good quantitative comparison of schools receiving intervention versus those receiving no intervention to determine effectiveness of the program. A study of this nature would provide a well rounded analysis of the SMP's ability to affect a specific area of focus. In other words, the study would analyze schools who determined Biology to be an area of focus and compare their results from the Biology End of Course Test to schools receiving no intervention. In addition, the action plans would be used as a basis to interview teachers to determine if they felt they made gains in additional areas such as percentage of time spent in inquiry or their placement on the spectrum of inquiry (Olson & Loucks-Horsley 2000).

Another very interesting aspect of improving science achievement in a state where scientific practices are required is to evaluate the resources available to treatment schools and non-treatment schools. School systems that have the ability to support their science teachers through training, central office support, and financial resources that allow for quality on-going laboratory experiences would most likely have an advantage over school systems that either do not or are not able to render this support to its schools. A per pupil cost analysis could render interesting results as could metro-Atlanta schools versus high-level intervention schools. Metro Atlanta schools tend to have central office support and the finances to support its teachers in a way that many rural systems do not. A study of this nature would have to begin with a state-level view of the distribution and allotment of state-funding and compare it to the amount of funding actually utilized in the science classrooms. This would be an intensive study as it would need to analyze department budgets and expenditures over several years to establish a trend.

Key factors to analyze would be the number of teachers and their class sizes, department budget requests versus actual allotments, and expenditures on lab equipment and supplies.

Finally, a qualitative analysis of teachers would be most critical to determine effectiveness at the classroom level. The qualitative analysis could evaluate the perception of the teachers and administrators who worked with the SMP. The research may be both survey and interview-based. The survey could be done in a way to evaluate the overall perception of the impact of the program. The survey could also include questions that focus on practices prior to, during and after the intervention. Particular attention could be given to awareness and implementation of strategies to enhance inquiry instruction and differentiate instruction as a whole for SWD. The questions may also focus on how often teachers conducted laboratory experiments, what was their perception of quality inquiry, and what was their understanding of the GPS and GHSGT. The purpose of the interview portion of the study could be to expand on the survey. While it would have the same general themes, it could focus on the teachers' and administrators' perception of how their classroom has changed as a result of the intervention especially with regard to differentiation strategies specifically designed to support SWD and ED. The interview should be conducted throughout the state with quality representation from each SMP region. Another interesting and potentially informative round of interviews could be with science staff from the Regional Education Service Agencies (RESA) to assess their perception of the success of the program. The RESA staff did not embrace the SMP at its inception, but over time the SIS and GaDOE staff felt they learned to work well together for a common goal. If possible, an analysis of the pre-intervention and post-intervention EOCT compared to interviewed teachers could provide additional evidence as to the effectiveness of the program at a more granular level.

Operational Definitions

Constructivism is the belief that students construct their own knowledge.

Contextualized scientific practice is the practice of assessing scientific practices (i.e. Habits of Mind/Nature of Science in the context of content for the purpose of assessing a student's ability to apply knowledge to new situations.

End-of-course test (EOCT) is the Georgia assessment given at the end of high school Biology and Physical Science courses that meets House Bill 1187 Georgia General Assembly legislation.

Factoid refers to test items that do not exceed the knowledge level of Bloom's Taxonomy. Factoids are also generally considered trivial pieces of information. (Joiner, 2004)

Georgia High School Graduation Test (GHSGT) is the Georgia assessment consisting of life science and physical science that meets *No Child Left Behind* legislation.

Hands-on Instruction is often times used interchangeably with inquiry learning. For this paper it is used as a tool for inquiry learning.

House Bill 1187 was passed in 2000 as the education reform law championed by the Governor of Georgia, Roy Barnes.

Inquiry Learning refers to the activities of students in which they develop knowledge and understanding of scientific ideas, as well as an understanding of how scientists study the natural world. (NSES, p. 2)

Pedagogical Content Knowledge is the ability of the teacher to assess and redirect instruction to counter a student's misconception of a scientific idea.

Proficiency Rate is defined as the percent of students meeting or exceeding the designated cut score on the science portion of the GHSGT. During the time of this study, students had to score 500 on the GHSGT to meet the standard.

Science Implementation Specialists (SIS) are GaDOE staff hired for the specific purpose of mentoring and coaching struggling science teachers in the areas of content and inquiry pedagogy.

Scientific Practices refer to the tools used in inquiry learning such as items defined in *Benchmarks for Science Literacy* under the sections for Habits of Mind and Nature of Science.

Scientific Literacy is the knowledge and understanding of scientific concepts and processes required for personal decision making, participation in civic and cultural affairs, and economic productivity. It also includes specific types of abilities. (NSES, p.1)

Student Achievement refers, for the purpose of this study, is indicated by the average scale score on the science portion of the GHSGT.

References

Aguilar, J. (2009). Personal Communication.

Alparson, C., Tekkaya, C., & Geban O. (2003). Using the conceptual change instruction to improve learning. *Journal of Biological Education*, 37(3), 133-137.

American Association for the Advancement of Science (AAAS). (1989). Science for all Americans: A project 2061 report on literacy goals in science, mathematics, and technology. Washington DC: American Association for the Advancement of Science.

Applefield, J., Huber, R., & Moallem, M. (2000/2001). Constructivism in theory and practice: toward a better understanding. *High School Journal*, 84 (2), 35-54.

Atwood, R.K.&Oldham, B.R. (1985). Teachers' perceptions of mainstreaming in an Inquiry-oriented elementary science program. *Science Education*, 69 (5), 619–624.

Ball, D., & Cohen, D. (1999). Developing practice, developing practioners: toward a practice-based theory of professional education. In G. Sykes & L. Darling-Hammond (Eds.), *Teaching as the learning profession: Handbook of policy and practice* (pp. 3-32). San Francisco, Ca: Jossey Bass.

Barton, A., & Tan, E. (2009). Funds of knowledge and discourses and hybrid space. *Journal of Research in Science Teaching*, *46*(1), 50–73.

Basu, S., & Barton, A. (2007). Developing a sustained interest in science among urban minority youth. *Journal of Research in Science Teaching*, *44*(3), 466–489.

Beaumont-Walters, Y., & Soyibo, K. (2001). An Analysis of High School Students' Performance on Five Integrated Science Process Skills. *Research in Science & Technological Education*, 19 (2), 133-26.

Boyd, G. (2004). Retrieved Mar 28, 2004

Bransford, Brown, and Cocking (NRC, 2000b)

Bredderman, T. (1982). The effects of activity-based science in elementary schools. In M. Rowe & W. Higuchi (Eds.), Education in the 80's (pp.63-75). Washington, DC: National Education Association.

Brophy, J. (1992). Probing the subtleties of subject-matter teaching. *Educational Leadership,* 49 (7), 4-8.

Bybee, R. (2009). Program for international student assessment (pisa) 2006 and scientific literacy: a perspective for science education leaders. *Science Educator, 18*(2), 1-14.

Cawley, J., Foley, T., & Miller, J. (2003). Science and students with mild disabilities. *Intervention in School & Clinic*, 38 (3), 160-171.

Chin, C., & Brown, D. (2002). Student-generated questions: a meaningful aspect of learning in science. *International Journal of Science Education*, 24 (5), 521-550.

Council for State Science Supervisors, (n.d.). Big picture overview of NLIST vision. retrieved Feb 15, 2004, from Networking for Leadership, Inquiry and Systemic Thinking Web site: http://www.inquiryscience.com.

Dalton, B., & Morocco, C. C. (1997). Supported inquiry science: Teaching for conceptual change in urban and suburban science classrooms. Journal of Learning Disabilities, 30, 670–685.

Danielson, C. (1996). *Enhancing Professional Practice: A Framework for Teaching.* Alexandria, VA: Association for Supervision and Curriculum Development.

Darling-Hammond, L. (2004). *Inequality and the right to learn: Access to qualified teachers in California's public schools*. Teacher's College Record, 106, 1936–1966.

Dossey, J.A., McCrone, S.A., and O'Sullivan, C. (2006). *Problem Solving in the PISA and TIMSS 2003 Assessments* (NCES 2007-049). U.S. Department of Education. Washington, DC: National Center for Education Statistics. Retrieved August 25, 2009 from http://nces.ed.gov/pubsearch.

Driver, R., Asoko, H., Leach, J., Mortimer, E., & Scott, P. (1994). Constructing scientific knowledge in the classroom. *Educational Researcher,* 23, 5–12.

Driver, R. (1989). The construction of scientific knowledge in school classrooms. In R. Miller (Ed.), *Doing science: images of science in science education* (pp. 83–106). Bristol, PA: Taylor & Francis. Retrieved May 23, 2010 from http://www.eric.ed.gov/ERICDocs/data/ericdocs2sql/content_storage_01/0000019b/80/1f/47/d1.pdf.

Fincher, M. (2009). Personal communication.

Georgia Department of Education, (2006). Released GHSGT results.

Georgia Department of Education, (2005). Released GHSGT results.

Georgia Department of Education, (2005). Science Mentor Program Handbook.

Georgia Department of Education, (2004). GPS Public review survey.

Georgia Department of Education (2003). Science staff review of state standards.

Georgia Department of Education (2002). GPS Phase I Task Force.

Georgia Department of Education, (2002). Study on the effect of block schedule on student performance on the Georgia High School Graduation Test.

Georgia State Board of Education (2004). State Board resolution on expectations of the Georgia Department of Education to revise Georgia's standards.

Geier, R., Blumenfeld, P., Marx, R., Krajcik, J., Fishman, B., Solloway, E., & Clay-Chambers, J. (2008). Standardized test outcomes for students engaged in inquiry-based science curricula in the context of urban reform. *Journal of Research in Science Teaching*, *45*(8), 922–939.

Grigg, W., Lauko, M., & Brockway, D. (2006, May 24). *The Nation's report card: science 2005*. Retrieved from http://nationsreportcard.gov/science_2005/

Gruber, H.E. & Voneche, J.J. (1977). The essential Piaget. New York: Basic Books.

Gurganus, S., Janas, M., & Schmitt, L. (1995). Science instruction: What special education teachers need to know and what roles they need to play. *Teaching Exceptional Children*, 27 (4), 7–9.

Hodson, D. (1996). Laboratory work as scientific method: three decades of confusion and distortion, *Journal of Curriculum Studies*, 28, pp. 115-135.

Huffman, D., Thomas, K. & Lawrenz, F. (2003) Relationship between professional development, teachers' instructional practices, and the achievement of students in science and mathematics. *School Science and Mathematics*, 103 (8), 378-387.

Huck, S.W. (2000). *Reading statistics and research*. New York, NY: Addison Wesley Longman.

Kanevsky, L., & Keighley, T. (2003). To Produce or Not to Produce? Understanding Boredom and the Honor in Underachievement, *Roeper Review*, 26 (1) 20-29.

Kent, A.M., Feldman, P., & Hayes, R.L. (2009). Mentoring and inducting new teachers into the profession: an innovative approach. *International Journal of Applied Educational Studies*, 5(1), Retrieved from http://www.ijaes.com/archive/2009/volume5/Abstract5-1-6.pdf

Kohn, A. (1993). Choices for children: Why and how to let students decide. *Phi Delta Kappan*, 75 (1), 8-16

Klum, G., & Stuessy, C. (1991). Assessment in science and mathematics education reform (Chapter 5). In G. Klum and S. Malcom (Eds.), *Science assessment in the service of reform.* Washington, DC: American Association for the Advancement of Science.

Larson, R. (1989). Beeping children and adolescents: A method for studying time use and daily experience. *Journal of Youth and Adolescence*, 18 (6), 511-530.

Lee, O., Buxton, C., Lewis, S., & LeRoy, K. (2006). Science inquiry and student diversity: enhanced abilities and continuing difficulties after an instructional intervention. *Journal of Research in Science Teaching, 43*(7), 607-636.

Lee, O., & Luykx, A. (2005). Dilemmas in scaling up educational innovations with nonmainstream students in elementary school science. *American Educational Research Journal*, 43, 411–438.

Lee, O., & Paik, S. (2000). Conceptions of science achievement in major reform documents. *School Science & Mathematics*, *100*(1), 16-26.

Llewellyn, D. (2002). *Inquire within: implementing inquiry-based science standards*. Thousand Oaks, Ca: Corwin Press, Inc.

Penfield, R., & Okee, O. (2010). Test-based accountability: potential benefits and pitfalls of science assessment with student diversity. *Journal of Research in Science Teaching*, *47*(1), 6-24.

Leedy, P. D., & Ormrod, J. E. (2005). *Practical research* (8th ed.). Upper Saddle River, NJ: Pearson Merrill Prentice Hall.

Mastropieri, M, & Scruggs, T. (2006). Differentiated curriculum enhancement in inclusive middle school science: effects on classroom and high-stakes tests. *Journal of Special Education*, *40*(3), 130-137.

Mastropieri, M. A., Scruggs, T. E., Mantzicopoulos, P., Sturgeon, A., Goodwin, L., et al. (1998). "A place where living things affect and depend on each other": Qualitative and quantitative outcomes associated with inclusive science teaching. *Science Education*, *82*, 163–180.

Massachusetts Department of Education, (2005). MCAS released test items 2005.

McCarthy, C. B. (2005). Effects of thematic-based, hands-on science teaching versus a textbook approach for students with disabilities. *Journal of Research in Science Teaching*, *42*(3), 245-263.

Mehalik, M., Doppelt, Y., & Schuun, C. (2008). Middle-school science through design-based learning versus scripted inquiry: better overall science concept learning and equity gap reduction. *Journal of Engineering Education*, *97*(1), 71-85.

Melville, W., & Bartley, A. (2010). Mentoring and community: inquiry as stance and science as inquiry. *International Journal of Science Education*, *32*(6), 807-828.

National Assessment of Educational Progress, Initials. (2009, October 22). *Grade 8: using the floating pencil test to estimate salt concentration of an unknown salt solution.*. Retrieved from http://nationsreportcard.gov/science_2005/s0116.asp

National Center for Educational Statistics, (2000). National assessment of educational progress. Retrieved Mar 21, 2004, from http://nces.ed.gov.nationsreportcard/.

National Center for Educational Statistics, Third International Mathematics and Science Study. 1995. Retrieved October 15, 2003, from http://nces.ed.gov.timss.

National Research Council (1996). *National Science Education Standards*. Washington, DC. National Academy Press.

Neidorf, T.S., Binkley, M., and Stephens, M. (forthcoming). *Comparing Science Content in the National Assessment of Educational Progress (NAEP) 2000 and Trends in International Mathematics and Science Study (TIMSS) 2003 Assessments* (NCES 2006–026). U.S. Department of Education. Washington, DC: National Center for Education Statistics. Retrieved August 22, 2009 from http://nces.ed.gov/pubsearch.

Olson S. and Loucks-Horsley S. (2000). *Inquiry and the National Science Education Standards*, Washington DC. National Academy Press.

Page, Rod. "Opening Remarks to Science Supervisors." Science Education Summit. , Washington DC. 15 March 2004.

Peressini, D., Borko, H., Romagnano, L., Knuth, E., & Willis, C. (2004). A conceptual framework for learning to teach secondary mathematics: a situative perspective. *Educational Studies in Mathematics*, *56*(1), 67-96.

Rhem, J. (2003). Mindfulness and teaching. *The National Teaching and Learning FORUM, (1) 1-3.*

Rutherford, F. J., & Ahlgren, A. (1990). *Project 2061*. New York, NY: Oxford University Press.

Rutherford, F. J., & Ahlgren, A. (1990). *Science for all Americans*. New York, NY: Oxford University Press.

Saurino, D., & Saurino, P. (1999). Making efficient use of mentoring programs: a collaborative group action research approach. *Proceedings of the Annual meeting of the national association for research in science teaching* (pp. 1-18). Boston: National Association for Research in Science Teaching.

Seymour, E.& Hewitt, NM. (1997). Talking about leaving. Why undergraduates leave the sciences. Westview Press: Boulder, CO.

Shymansky, J., Kyle, W. Jr., & Alport, J. (1983). The effects of new science curricula on student performance, *Journal of Research in Science,* 20(5), 387-404.

Stohr-Hunt, P. (1996). An analysis of frequency of hands-on experience and science achievement. *Journal of Research in Science Teaching*. 33 (1), 101-109.

Tippins, D., & Tobin, K. (1993). Ethical decisions at the heart of teaching: Making sense from a constructivist perspective. *Journal of Moral Education*, *22*(3), 221. Retrieved from Academic Search Premier database.

Tretter, T., & Jones, M. (2003). Relationships between inquiry-based teaching and physical science standardized test scores. *School* Science *& Mathematics*, 103 (7), 345-351.

University System of Georgia (2003). Partnership for Reform in Mathematics and Science Education Grant Application. *National Science Foundation.*

Veal, W., & Flinders, D. (2001). How Block Scheduling Reform Effects Classroom Practice. *High School Journal*, 84. Retrieved Feb 16, 2004, from http://search.epnet.com/direct.asp?an-4390127&db=afh.

von Secker, C. (2002). Effects of inquiry-based teacher practices on science excellence and equity. *The Journal of Educational Research*, 95 (3), 151-160.

Warren, B., Ballenger, C., Ogonowski, M., Rosebery, A., & Hudicourt-Barnes, J. (2001). Rethinking diversity in learning science: the logic of everyday sense-making. *Journal of Research in Science Teaching*, *38*(5), 529-552.

Watts, M., Gould, G., & Alsop, S. (1997). Questions of understanding: categorizing pupil’s questions in science. *School Science Review*, 79 (286), 57-63.

Weiss, I., & Pasley, J. (2004). What is high-quality instruction? *Educational Leadership*, 61 (5), 24-29.

Weiss, I.R., Pasley, J.D., Smith, P.S., Banilower, E.R.,&Heck, D.J. (2003). Looking inside the classroom: a study of K-12 mathematics and science education in the U.S. Chapel Hill, NC: Horizon Research.

Wheeler, J. (2004). National science teacher association update. *Council for State Science Supervisors*, Retrieved Mar 28, 2004.

Wilson, C., Taylor, J., Kowalski, S., & Carlson, J. (2010). The Relative effects and equity of inquiry-based and commonplace science teaching on students' knowledge, reasoning, and argumentation. *Journal of Research in Science Teaching*, *47*(3), 276-301.

Yager, R. (1991). The constructivist learning model. *Science Teacher*, 58, 52-57.

Wilson, M. & Bertenthal, M (2005). *Systems for State Science Assessment*. Washington, D.C.: The National Academies Press.

Zelditch, M. (1990). "Mentor roles," in Proceedings of the 32nd Annual Meeting of the Western Association of Graduate Schools, 11. Tempe, Ariz., March 16-18.).

Zembylas, M., & Isenbarger, L. (2002). Teaching science to students with learning disabilities: subverting the myths of labeling through teachers' caring and enthusiasm. *Research in Science Education*, *32* (1), 55-79.

Appendices

Appendix A

List of Schools Receiving High-Level Intervention by the SMP

Schools	**School System**	**Intervention Level**
Atkinson County High School	Atkinson County	High
Bacon County High School	Bacon County	High
Baldwin County High School	Baldwin County	High
Brantley County High School	Brantley County	High
Brooks County High School	Brooks County	High
Cairo High School	Grady County	High
Carver High School	Muscogee County	High
Cedartown High School	Polk County	High
Central High School	Talbot County	High
Chattooga High School	Chattooga County	High
Chestatee High School	Hall County	High
Colquitt County High School	Colquitt County	High
Columbia High School	Dekalb County	High
Cross Keys High School	Dekalb County	High
Dodge County High School	Dodge County	High
Dooley County High School	Dooley County	High
Early County High School	Early County	High
Fitzgerald High School	Ben Hill County	High
Franklin County High School	Franklin County	High
Glascock County High School	Glascock County	High
Greenville High School	Meriwether County	High
Griffin High School	Spalding County	High
Irwin County High School	Irwin County	High
Jackson High School	Butts County	High
Jefferson County High School	Jefferson County	High
Kendrick High School	Muscogee County	High
Lanier County High School	Lanier County	High
Lithia Springs Comp. High	Douglas County	High
Lowndes County High School	Lowndes County	High
Madison County High School	Madison County	High
Manchester High School	Meriwether County	High
McIntosh County Academy	McIntosh County	High
Mitchell County High School	Mitchell County	High
Murray County High School	Murray County	High
Oglethorpe County High School	Oglethorpe County	High
Peach County High School	Peach County	High
Ridgeland High School	Walker County	High
Seminole County High School	Seminole County	High

List of Schools Receiving High-Level Intervention by the SMP (cont)

Schools	**School System**	**Intervention Level**
Stewart-Quitman High School	Stewart County	High
Taliaferro County High School	Taliaferro County	High
Telfair County High School	Telfair County	High
Terrell County High School	Terrell County	High
Thomasville High School	Thomasville City Schools	High
Turner County High School	Turner County	High
Valdosta High School	Valdosta City Schools	High
Villa Rica High School	Carroll County	High
Warren County High School	Warren County	High
Wilkinson County High School	Wilkinson County	High
Worth County High School	Worth County	High

Appendix B

List of Schools Receiving Medium-Level Intervention by the SMP

Schools	**School System**	**Intervention Level**
Americus High School South	Sumter County	Medium
Bradwell Institute	Liberty County	Medium
Burke County High School	Burke County	Medium
Charlton County High School	Charlton County	Medium
Clarke Central High School	Clarke County	Medium
Clinch County High School	Clinch County	Medium
Coffee County High School	Coffee County	Medium
Creekside High School	Fulton County	Medium
Dougherty Comp. High School	Dougherty County	Medium
East Hall High School	Hall County	Medium
Glenn Hills High School	Richmond County	Medium
Hancock Central High School	Hancock County	Medium
Haralson County High School	Haralson County	Medium
Hephzibah High School	Richmond County	Medium
Jasper County High School	Jasper County	Medium
Lafayette High School	Walker County	Medium
Lithonia High School	Dekalb County	Medium
McNair High School	Dekalb County	Medium
MLK High School	Dekalb County	Medium
Paulding County High School	Paulding County	Medium
Upson-Lee High School	Thomaston-Upson County	Medium
Ware County High School	Ware County	Medium

Appendix C

List of Comparison Schools Receiving No Intervention by the SMP

Schools	**School System**	**Similar Group A School**	**Similarity Index**
Echols County High School	Echols County	Atkinson County High School	3
Wayne County High School	Wayne County	Bacon County High School	1
Washington County High School	Washington County	Baldwin County High School	1
Rabun County High School	Rabun County	Brantley County High School	1
Jordan Vocational School	Muscogee County	Brooks County High School	1
Rutland High School	Bibb County	Cairo High School	1
Washington High School	Atlanta Public	Carver High School	1
Bleckley County High School	Bleckley County	Cedartown High School	1
Monroe High School	Dougherty County	Central High School	1
Mt. Zion High School	Carroll County	Chattooga High School	1
Calhoun High School	Calhoun City	Chestatee High School	2
Long County High School	Long County	Colquitt County High School	1
Miller Grove High School	Dekalb County	Columbia High School	1
No Similar School		Cross Keys High School	
Emmanuel Institute	Emanuel County	Dodge County High School	1
Randolph County High School	Randolph County	Dooley County High School	2
Washington-Wilkes High School	Wilkes County	Early County High School	1
Wilcox County High School	Wilcox County	Fitzgerald High School	1
Pierce County High School	Pierce County	Franklin County High School	1
East Jackson High School	Jackson County	Glascock County High School	1
Twiggs County High School	Twiggs County	Greenville High School	1
Taylor County High School	Taylor County	Griffin High School	2
Treutlen High School	Treutlen County	Irwin County High School	1
Mary Persons High School	Monroe County	Jackson High School	1

List of Comparison Schools Receiving No Intervention by the SMP (cont)

Schools	**School System**	**Similar Group A School**	**Similarity Index**
Greene County High School	Greene County	Jefferson County High School	1
Douglass High School	Atlanta Public	Kendrick High School	1
Portal Middle/High School	Bullock County	Lanier County High School	1
Callaway High School	Troup County	Lithia Springs Comp. High	2
Effingham County High School	Effingham County	Lowndes County High School	1
Stephens County High School	Stephens County	Madison County High School	1
Thomson High School	McDuffie County	Manchester High School	1
Claxton High School	Evans County	McIntosh County Academy	1
Southwest High School	Bibb County	Mitchell County High School	2
Banks County High School	Banks County	Murray County High School	2
Commerce High School	Commerce City	Oglethorpe County HS	1
Crisp County High School	Crisp County	Peach County High School	1
Gordon Central High School	Gordon County	Ridgeland High School	2
Swainsboro High School	Emanuel County	Seminole County High School	1
Josey High School	Richmond County	Stewart-Quitman High School	1
Spencer High School	Muscogee County	Taliaferro County High School	6
Johnson County High School	Johnson County	Telfair County High School	1
Stephenson High School	Dekalb County	Terrell County High School	1
Butler High School	Richmond County	Thomasville High School	1
Jenkins County High School	Jenkins County	Turner County High School	1
Academy of Richmond County	Richmond County	Valdosta High School	1
Appling County High School	Appling County	Villa Rica High School	1
School of Technology at Carver	Atlanta Public	Warren County High School	2
Ceder Shoals High School	Clarke County	Wilkinson County High School	3
LaGrange High School	Troup County	Worth County High School	4

Appendix D

All Student Scale Score Analysis

Descriptive Statistics

	Treatment Group	Mean	Std. Deviation	N
2004 GHSGT All Students	High-Level Intervention	502.04	6.522	47
	Medium-Level Intervention	504.09	4.657	22
	No Intervention	504.36	6.182	45
	Total	503.35	6.120	114
2005 GHSGT All Students	High-Level Intervention	501.80	7.100	47
	Medium-Level Intervention	501.40	5.586	22
	No Intervention	502.72	6.917	45
	Total	502.09	6.725	114
2006 GHSGT All Students	High-Level Intervention	504.96	6.813	47
	Medium-Level Intervention	505.09	5.580	22
	No Intervention	505.86	6.680	45
	Total	505.34	6.499	114
2007 GHSGT All Students	High-Level Intervention	506.08	6.385	47
	Medium-Level Intervention	505.15	4.851	22
	No Intervention	507.65	7.238	45
	Total	506.52	6.505	114

Appendix E

All Student Proficiency Analysis

Descriptive Statistics

	Treatment Group	Mean	Std. Deviation	N
2004 GHSGT All Students	High-Level Intervention	54.1975	14.63397	48
	Medium-Level Intervention	58.4208	8.75402	22
	No Intervention	58.2494	13.31321	45
	Total	56.5910	13.22727	115
2005 GHSGT All Students	High-Level Intervention	56.0012	11.79446	48
	Medium-Level Intervention	53.7334	11.64757	22
	No Intervention	57.5589	12.47360	45
	Total	56.1769	12.01311	115
2006 GHSGT All Students	High-Level Intervention	61.3418	13.09121	48
	Medium-Level Intervention	60.8302	12.10335	22
	No Intervention	63.9668	13.33988	45
	Total	62.2711	12.97013	115
2007 GHSGT All Students	High-Level Intervention	64.1246	12.30704	48
	Medium-Level Intervention	63.2090	7.60850	22
	No Intervention	68.1152	10.02984	45
	Total	65.5110	10.79099	115

Appendix F
Gender Scale Score Analysis

Descriptive Statistics

	Treatment Group	Subgroups	Mean	Std. Deviation	N
2004 GHSGT	High-Level Intervention	Male	504.17	7.912	48
		Female	500.21	5.686	48
		Total	502.19	7.136	96
	Medium-Level Intervention	Male	506.43	5.864	22
		Female	502.03	4.198	22
		Total	504.23	5.509	44
	No Intervention	Male	506.57	7.099	45
		Female	502.14	5.797	45
		Total	504.36	6.819	90
	Total	Male	505.54	7.278	115
		Female	501.32	5.516	115
		Total	503.43	6.782	230
2005 GHSGT	High-Level Intervention	Male	503.62	8.618	48
		Female	500.38	6.528	48
		Total	502.00	7.777	96
	Medium-Level Intervention	Male	503.13	7.381	22
		Female	499.84	5.128	22
		Total	501.49	6.497	44
	No Intervention	Male	504.74	7.823	45
		Female	500.98	6.806	45
		Total	502.86	7.532	90
	Total	Male	503.96	8.044	115
		Female	500.51	6.362	115
		Total	502.24	7.440	230

Gender Scale Score Analysis
Descriptive Statistics (cont)

2006 GHSGT	High-Level Intervention	Male	506.94	8.411	48
		Female	503.09	6.138	48
		Total	505.02	7.575	96
	Medium-Level Intervention	Male	507.25	5.843	22
		Female	503.26	5.573	22
		Total	505.25	5.993	44
	No Intervention	Male	506.79	8.219	45
		Female	505.07	5.876	45
		Total	505.93	7.157	90
	Total	Male	506.94	7.846	115
		Female	503.90	5.956	115
		Total	505.42	7.116	230
2007 GHSGT	High-Level Intervention	Male	508.70	7.690	48
		Female	503.88	5.921	48
		Total	506.29	7.244	96
	Medium-Level Intervention	Male	507.22	5.403	22
		Female	503.26	4.839	22
		Total	505.24	5.450	44
	No Intervention	Male	510.00	7.996	45
		Female	505.60	7.297	45
		Total	507.80	7.927	90
	Total	Male	508.93	7.448	115
		Female	504.44	6.344	115
		Total	506.68	7.261	230

Appendix G

Gender Proficiency Analysis

Descriptive Statistics

	Treatment Group	Subgroups	Mean	Std. Deviation	N
2004 GHSGT Percent Proficiency	High-Level Intervention	Male	58.0507	16.88570	48
		Female	50.8227	14.07235	48
		Total	54.4367	15.88190	96
	Medium-Level Intervention	Male	63.4204	10.63719	22
		Female	54.0167	8.39794	22
		Total	58.7186	10.59827	44
	No Intervention	Male	62.9956	13.34139	45
		Female	53.7499	14.72887	45
		Total	58.3727	14.72611	90
	Total	Male	61.0129	14.61025	115
		Female	52.5792	13.43897	115
		Total	56.7961	14.62983	230
2005 GHSGT Percent Proficiency	High-Level Intervention	Male	60.0813	12.76540	48
		Female	52.4177	12.40977	48
		Total	56.2495	13.10145	96
	Medium-Level Intervention	Male	58.1490	13.29915	22
		Female	49.8827	11.12656	22
		Total	54.0158	12.81866	44
	No Intervention	Male	61.7046	12.50173	45
		Female	54.0708	14.12075	45
		Total	57.8877	13.80502	90
	Total	Male	60.3468	12.71918	115
		Female	52.5796	12.86661	115
		Total	56.4632	13.34531	230

Gender Proficiency Analysis
Descriptive Statistics (cont)

2006 GHSGT Percent Proficiency	High-Level Intervention	Male	64.7499	15.75245	48
		Female	58.2686	12.35132	48
		Total	61.5092	14.45167	96
	Medium-Level Intervention	Male	64.5074	11.81992	22
		Female	57.7454	13.53282	22
		Total	61.1264	13.01410	44
	No Intervention	Male	66.3115	14.41335	45
		Female	62.1112	13.61773	45
		Total	64.2113	14.10123	90
	Total	Male	65.3145	14.45247	115
		Female	59.6721	13.11868	115
		Total	62.4933	14.05876	230
2007 GHSGT Percent Proficiency	High-Level Intervention	Male	68.6885	13.26460	48
		Female	60.0136	12.74758	48
		Total	64.3511	13.65486	96
	Medium-Level Intervention	Male	66.1379	9.23812	22
		Female	60.4665	7.92196	22
		Total	63.3022	8.97531	44
	No Intervention	Male	70.9943	11.15519	45
		Female	65.6465	10.27840	45
		Total	68.3204	10.99906	90
	Total	Male	69.1028	11.80981	115
		Female	62.3045	11.25178	115
		Total	65.7036	12.00254	230

Appendix H
Ethnicity Scale Score Analysis
Descriptive Statistics

	Treatment Group	Subgroups	Mean	Std. Deviation	N
2004 GHSGT Scale Score	High-Level Intervention	White Students	509.11	9.034	46
		Black Students	494.51	5.083	45
		Hispanic Students	502.05	15.146	30
		Asian Students	510.09	19.135	24
		Total	503.28	13.454	145
	Medium-Level Intervention	White Students	506.68	10.430	20
		Black Students	499.42	5.505	22
		Hispanic Students	499.53	13.097	15
		Asian Students	512.48	13.033	12
		Total	503.82	11.336	69
	No Intervention	White Students	508.89	8.900	44
		Black Students	496.21	6.690	45
		Hispanic Students	496.93	15.753	34
		Asian Students	504.52	17.380	20
		Total	501.44	12.962	143
	Total	White Students	508.58	9.204	110
		Black Students	496.16	6.077	112
		Hispanic Students	499.37	15.048	79
		Asian Students	508.61	17.371	56
		Total	502.65	12.876	357

Ethnicity Scale Score Analysis
Descriptive Statistics (cont)

2005 GHSGT Scale Score	High-Level Intervention	White Students	510.28	7.064	46
		Black Students	495.40	9.186	45
		Hispanic Students	495.39	18.982	30
		Asian Students	509.54	17.571	24
		Total	502.46	14.735	145
	Medium-Level Intervention	White Students	507.14	12.623	20
		Black Students	494.80	5.928	22
		Hispanic Students	501.28	11.935	15
		Asian Students	514.86	14.204	12
		Total	503.27	13.021	69
	No Intervention	White Students	510.35	5.887	44
		Black Students	494.00	6.718	45
		Hispanic Students	497.74	37.378	34
		Asian Students	510.05	12.987	20
		Total	502.17	20.665	143
	Total	White Students	509.74	7.972	110
		Black Students	494.72	7.641	112
		Hispanic Students	497.52	27.479	79
		Asian Students	510.86	15.239	56
		Total	502.50	17.050	357

Ethnicity Scale Score Analysis
Descriptive Statistics (cont)

2006 GHSGT Scale Score	High-Level Intervention	White Students	513.21	8.373	46
		Black Students	498.70	6.476	45
		Hispanic Students	503.60	14.297	30
		Asian Students	516.04	13.675	24
		Total	507.19	12.475	145
	Medium-Level Intervention	White Students	512.68	6.442	20
		Black Students	500.17	6.500	22
		Hispanic Students	504.06	9.055	15
		Asian Students	511.89	23.502	12
		Total	506.68	12.707	69
	No Intervention	White Students	512.02	6.926	44
		Black Students	497.19	10.829	45
		Hispanic Students	506.20	12.500	34
		Asian Students	511.13	15.890	20
		Total	505.85	12.667	143
	Total	White Students	512.64	7.443	110
		Black Students	498.38	8.506	112
		Hispanic Students	504.81	12.583	79
		Asian Students	513.40	16.773	56
		Total	506.55	12.576	357

Ethnicity Scale Score Analysis
Descriptive Statistics (cont)

2007 GHSGT Scale Score	High-Level Intervention	White Students	514.93	6.842	46
		Black Students	499.60	7.826	45
		Hispanic Students	505.14	14.636	30
		Asian Students	512.34	14.887	24
		Total	507.72	12.412	145
	Medium-Level Intervention	White Students	509.07	13.946	20
		Black Students	500.89	4.640	22
		Hispanic Students	504.85	9.485	15
		Asian Students	508.65	21.869	12
		Total	505.47	13.014	69
	No Intervention	White Students	512.14	15.961	44
		Black Students	498.21	10.949	45
		Hispanic Students	509.16	12.644	34
		Asian Students	521.07	19.538	20
		Total	508.30	16.202	143
	Total	White Students	512.75	12.585	110
		Black Students	499.29	8.768	112
		Hispanic Students	506.82	12.948	79
		Asian Students	514.67	18.584	56
		Total	507.52	14.158	357

Appendix I

Ethnicity Proficiency Analysis

Descriptive Statistics

	Treatment Group	Subgroups	Mean	Std. Deviation	N
2004 GHSGT Percent Proficiency	High-Level Intervention	White Students	70.8305	15.51786	46
		Black Students	38.0967	12.67944	48
		Hispanic Students	55.1644	35.32707	30
		Asian Students	66.3844	40.84268	24
		Total	56.3176	28.69592	148
	Medium-Level Intervention	White Students	66.2225	20.54102	19
		Black Students	47.4561	12.69922	22
		Hispanic Students	45.7488	33.73966	15
		Asian Students	76.5873	31.40418	12
		Total	57.4638	26.73114	68
	No Intervention	White Students	68.9344	17.41835	44
		Black Students	40.9321	15.54707	45
		Hispanic Students	43.1070	37.72590	34
		Asian Students	62.1154	42.21143	20
		Total	53.0280	29.93154	143
	Total	White Students	69.2619	17.15144	109
		Black Students	40.9967	14.17308	115
		Hispanic Students	48.1874	36.07742	79
		Asian Students	67.0461	39.23328	56
		Total	55.2244	28.81981	359

Ethnicity Proficiency Analysis
Descriptive Statistics (cont)

2005 GHSGT Percent Proficiency	High-Level Intervention	White Students	70.4256	19.21737	46
		Black Students	41.1705	14.79624	48
		Hispanic Students	46.2495	34.68645	30
		Asian Students	67.5683	40.09711	24
		Total	55.5736	29.15408	148
	Medium-Level Intervention	White Students	69.8227	15.06860	19
		Black Students	40.8313	10.92716	22
		Hispanic Students	51.8921	28.61924	15
		Asian Students	81.2500	30.59284	12
		Total	58.5044	25.77926	68
	No Intervention	White Students	71.5069	13.75117	44
		Black Students	39.4423	15.63036	45
		Hispanic Students	56.5501	31.60833	34
		Asian Students	63.7500	38.16659	20
		Total	56.7756	27.00687	143
	Total	White Students	70.7570	16.35356	109
		Black Students	40.4294	14.39433	115
		Hispanic Students	51.7540	32.22904	79
		Asian Students	69.1364	37.50182	56
		Total	56.6075	27.63731	359

Ethnicity Proficiency Analysis
Descriptive Statistics (cont)

2006 GHSGT Percent Proficiency	High-Level Intervention	White Students	75.6989	13.67187	46
		Black Students	50.3743	16.51189	48
		Hispanic Students	53.7240	33.97983	30
		Asian Students	81.3131	31.91267	24
		Total	63.9416	26.56044	148
	Medium-Level Intervention	White Students	78.9393	12.56737	19
		Black Students	51.4380	12.43032	22
		Hispanic Students	62.4390	26.01930	15
		Asian Students	57.4459	36.31588	12
		Total	62.6091	23.87956	68
	No Intervention	White Students	73.1858	18.82850	44
		Black Students	50.0410	16.98647	45
		Hispanic Students	66.7986	32.71359	34
		Asian Students	67.5833	39.61612	20
		Total	63.6003	27.32902	143
	Total	White Students	75.2493	15.79782	109
		Black Students	50.4474	15.88997	115
		Hispanic Students	61.0058	32.23326	79
		Asian Students	71.2952	36.36583	56
		Total	63.5532	26.32242	359

Ethnicity Proficiency Analysis
Descriptive Statistics (cont)

2007 GHSGT Percent Proficiency	High-Level Intervention	White Students	79.9799	10.79052	46
		Black Students	52.1838	15.97059	48
		Hispanic Students	61.8774	30.61800	30
		Asian Students	78.8826	31.09222	24
		Total	67.1176	24.62085	148
	Medium-Level Intervention	White Students	76.6222	11.23011	19
		Black Students	54.4390	8.40588	22
		Hispanic Students	57.6373	24.28023	15
		Asian Students	75.3373	26.33715	12
		Total	65.0307	19.98246	68
	No Intervention	White Students	78.1114	12.91225	44
		Black Students	52.2905	13.75948	45
		Hispanic Students	73.5653	25.07562	34
		Asian Students	81.2500	33.31962	20
		Total	69.3440	23.32459	143
	Total	White Students	78.6404	11.72649	109
		Black Students	52.6570	13.85658	115
		Hispanic Students	66.1025	27.66414	79
		Asian Students	78.9684	30.51708	56
		Total	67.6091	23.28076	359

Appendix J

Economically Disadvantaged Scale Score Analysis

Descriptive Statistics

	Treatment Group	Subgroups	Mean	Std. Deviation	N
2004 GHSGT Scale Score	High-Level Intervention	Economically Disadvantaged	496.39	5.985	48
		Non-Economically Disadvantaged	508.75	4.704	38
		Total	501.85	8.219	86
	Medium-Level Intervention	Economically Disadvantaged	500.77	5.479	22
		Non-Economically Disadvantaged	507.13	4.884	19
		Total	503.72	6.066	41
	No Intervention	Economically Disadvantaged	499.22	5.726	45
		Non-Economically Disadvantaged	508.63	7.063	38
		Total	503.53	7.897	83
	Total	Economically Disadvantaged	498.33	6.001	115
		Non-Economically Disadvantaged	508.38	5.771	95
		Total	502.88	7.728	210
2005 GHSGT Scale Score	High-Level Intervention	Economically Disadvantaged	495.95	10.191	48
		Non-Economically Disadvantaged	508.58	6.617	38
		Total	501.53	10.783	86
	Medium-Level Intervention	Economically Disadvantaged	497.79	4.668	22
		Non-Economically Disadvantaged	504.39	8.795	19
		Total	500.85	7.572	41
	No Intervention	Economically Disadvantaged	497.31	6.365	45
		Non-Economically Disadvantaged	508.55	8.055	38
		Total	502.46	9.096	83
	Total	Economically Disadvantaged	496.84	7.941	115
		Non-Economically Disadvantaged	507.73	7.772	95
		Total	501.76	9.545	210

Economically Disadvantaged Scale Score Analysis
Descriptive Statistics (cont)

2006 GHSGT Scale Score	High-Level Intervention	Economically Disadvantaged	500.41	5.170	48
		Non-Economically Disadvantaged	512.67	4.815	38
		Total	505.83	7.897	86
	Medium-Level Intervention	Economically Disadvantaged	501.13	4.733	22
		Non-Economically Disadvantaged	509.19	7.525	19
		Total	504.87	7.334	41
	No Intervention	Economically Disadvantaged	501.09	6.216	45
		Non-Economically Disadvantaged	511.02	7.298	38
		Total	505.63	8.340	83
	Total	Economically Disadvantaged	500.81	5.493	115
		Non-Economically Disadvantaged	511.31	6.529	95
		Total	505.56	7.942	210
2007 GHSGT Scale Score	High-Level Intervention	Economically Disadvantaged	501.51	5.724	48
		Non-Economically Disadvantaged	513.15	4.672	38
		Total	506.66	7.838	86
	Medium-Level Intervention	Economically Disadvantaged	502.22	4.844	22
		Non-Economically Disadvantaged	508.93	6.023	19
		Total	505.33	6.335	41
	No Intervention	Economically Disadvantaged	502.97	6.635	45
		Non-Economically Disadvantaged	513.07	9.236	38
		Total	507.59	9.365	83
	Total	Economically Disadvantaged	502.22	5.938	115
		Non-Economically Disadvantaged	512.27	7.207	95
		Total	506.77	8.231	210

Appendix K
Economically Disadvantaged Proficiency Analysis
Descriptive Statistics

	Treatment Group	Subgroups	Mean	Std. Deviation	N
2004 GHSGT Percent Proficient	High-Level Intervention	Economically Disadvantaged	42.2639	12.54204	48
		Non-Economically Disadvantaged	68.5533	10.38338	38
		Total	53.8801	17.50321	86
	Medium-Level Intervention	Economically Disadvantaged	51.0449	10.39648	22
		Non-Economically Disadvantaged	64.9983	9.54363	19
		Total	57.5111	12.13902	41
	No Intervention	Economically Disadvantaged	47.6970	13.17579	45
		Non-Economically Disadvantaged	67.7142	14.56194	38
		Total	56.8615	17.01485	83
	Total	Economically Disadvantaged	46.0698	12.79408	115
		Non-Economically Disadvantaged	67.5066	12.04480	95
		Total	55.7674	16.39885	210
2005 GHSGT Percent Proficient	High-Level Intervention	Economically Disadvantaged	47.0287	11.17300	48
		Non-Economically Disadvantaged	69.0744	10.08859	38
		Total	56.7698	15.31671	86
	Medium-Level Intervention	Economically Disadvantaged	46.5855	10.30430	22
		Non-Economically Disadvantaged	60.4231	14.90705	19
		Total	52.9981	14.30199	41
	No Intervention	Economically Disadvantaged	47.7245	9.87551	45
		Non-Economically Disadvantaged	68.3792	14.58657	38
		Total	57.1809	15.98510	83
	Total	Economically Disadvantaged	47.2162	10.43368	115
		Non-Economically Disadvantaged	67.0661	13.32708	95
		Total	56.1959	15.40584	210

Economically Disadvantaged Proficiency Analysis
Descriptive Statistics (cont)

2006 GHSGT Percent Proficient	High-Level Intervention	Economically Disadvantaged	52.9224	11.38524	48
		Non-Economically Disadvantaged	75.7438	8.49781	38
		Total	63.0063	15.26639	86
	Medium-Level Intervention	Economically Disadvantaged	53.2572	11.14320	22
		Non-Economically Disadvantaged	68.7588	15.08491	19
		Total	60.4409	15.12733	41
	No Intervention	Economically Disadvantaged	55.1939	12.96430	45
		Non-Economically Disadvantaged	73.3369	15.63509	38
		Total	63.5003	16.82830	83
	Total	Economically Disadvantaged	53.8753	11.93012	115
		Non-Economically Disadvantaged	73.3840	13.22107	95
		Total	62.7007	15.84252	210
2007 GHSGT Percent Proficient	High-Level Intervention	Economically Disadvantaged	56.6645	11.47914	48
		Non-Economically Disadvantaged	75.7147	7.83000	38
		Total	65.0820	13.78775	86
	Medium-Level Intervention	Economically Disadvantaged	57.2659	7.73747	22
		Non-Economically Disadvantaged	69.4151	12.73392	19
		Total	62.8960	11.91725	41
	No Intervention	Economically Disadvantaged	60.3965	10.24873	45
		Non-Economically Disadvantaged	77.2565	11.60670	38
		Total	68.1155	13.73198	83
	Total	Economically Disadvantaged	58.2399	10.43835	115
		Non-Economically Disadvantaged	75.0715	10.80617	95
		Total	65.8542	13.50819	210

Appendix L

Students With Disabilities Scale Score Analysis

Descriptive Statistics

	Treatment Group	Subgroups	Mean	Std. Deviation	N
2004 GHSGT Scale Scores	High-Level Intervention	SWD	481.95	8.208	47
		Non-SWD	504.30	6.445	48
		Total	493.24	13.412	95
	Medium-Level Intervention	SWD	483.52	8.700	22
		Non-SWD	505.75	4.238	22
		Total	494.63	13.120	44
	No Intervention	SWD	483.31	11.302	45
		Non-SWD	506.14	6.251	45
		Total	494.72	14.636	90
	Total	SWD	482.79	9.578	114
		Non-SWD	505.30	6.021	115
		Total	494.09	13.811	229
2005 GHSGT Scale Scores	High-Level Intervention	SWD	476.85	21.390	47
		Non-SWD	504.52	6.884	48
		Total	490.83	21.002	95
	Medium-Level Intervention	SWD	474.41	15.386	22
		Non-SWD	504.05	5.828	22
		Total	489.23	18.891	44
	No Intervention	SWD	474.88	28.978	45
		Non-SWD	505.04	6.886	45
		Total	489.96	25.857	90
	Total	SWD	475.60	23.629	114
		Non-SWD	504.63	6.650	115
		Total	490.18	22.594	229

Students With Disabilities Scale Score Analysis

Descriptive Statistics (cont)

2006 GHSGT Scale Scores	High-Level Intervention	SWD	483.11	17.789	47
		Non-SWD	507.30	6.343	48
		Total	495.33	17.965	95
	Medium-Level Intervention	SWD	483.43	7.378	22
		Non-SWD	507.09	5.925	22
		Total	495.26	13.670	44
	No Intervention	SWD	484.91	14.254	45
		Non-SWD	508.21	6.454	45
		Total	496.56	16.069	90
	Total	SWD	483.88	14.791	114
		Non-SWD	507.61	6.274	115
		Total	495.80	16.416	229
2007 GHSGT Scale Scores	High-Level Intervention	SWD	481.45	22.597	47
		Non-SWD	508.76	6.470	48
		Total	495.25	21.429	95
	Medium-Level Intervention	SWD	479.08	15.440	22
		Non-SWD	507.53	4.752	22
		Total	493.31	18.291	44
	No Intervention	SWD	481.19	28.809	45
		Non-SWD	510.85	5.927	45
		Total	496.02	25.496	90
	Total	SWD	480.89	24.003	114
		Non-SWD	509.34	6.053	115
		Total	495.18	22.519	229

Appendix M

Students With Disabilities Proficiency Analysis

Descriptive Statistics

	Treatment Group	Subgroups	Mean	Std. Deviation	N
2004 GHSGT Percent Proficiency	High-Level Intervention	SWD	17.6522	14.55374	47
		Non-SWD	58.2094	15.13300	48
		Total	38.1443	25.17353	95
	Medium-Level Intervention	SWD	21.2899	16.91691	22
		Non-SWD	61.3302	8.15887	22
		Total	41.3101	24.13301	44
	No Intervention	SWD	21.2899	16.91691	22
		Non-SWD	61.3133	13.40036	46
		Total	48.3646	23.79332	68
	Total	SWD	19.4111	15.65733	91
		Non-SWD	60.0322	13.35598	116
		Total	42.1746	24.80246	207
2005 GHSGT Percent Proficiency	High-Level Intervention	SWD	16.8809	14.61735	47
		Non-SWD	60.0464	11.97841	48
		Total	38.6908	25.43655	95
	Medium-Level Intervention	SWD	15.0757	12.55023	22
		Non-SWD	57.4356	12.59709	22
		Total	36.2557	24.76783	44
	No Intervention	SWD	15.0757	12.55023	22
		Non-SWD	61.1038	12.68434	46
		Total	46.2123	25.06034	68
	Total	SWD	16.0081	13.54748	91
		Non-SWD	59.9705	12.34186	116
		Total	40.6440	25.37040	207

Students With Disabilities Proficiency Analysis
Descriptive Statistics (cont)

2006 GHSGT Percent Proficiency	High-Level Intervention	SWD	23.2320	20.36720	47
		Non-SWD	65.5683	12.65614	48
		Total	44.6229	27.12739	95
	Medium-Level Intervention	SWD	17.9376	14.73740	22
		Non-SWD	64.7796	12.86897	22
		Total	41.3586	27.35413	44
	No Intervention	SWD	17.9376	14.73740	22
		Non-SWD	68.4036	13.24508	46
		Total	52.0764	27.41556	68
	Total	SWD	20.6721	17.90123	91
		Non-SWD	66.5431	12.91237	116
		Total	46.3776	27.45837	207
2007 GHSGT Percent Proficiency	High-Level Intervention	SWD	27.3144	20.14889	47
		Non-SWD	68.3371	12.72581	48
		Total	48.0416	26.54782	95
	Medium-Level Intervention	SWD	16.8121	14.45443	22
		Non-SWD	67.4229	7.99360	22
		Total	42.1175	28.08020	44
	No Intervention	SWD	16.8121	14.45443	22
		Non-SWD	72.5302	9.53157	46
		Total	54.5037	28.56732	68
	Total	SWD	22.2364	18.24421	91
		Non-SWD	69.8265	10.87936	116
		Total	48.9052	27.78759	207

Appendix N

Chronology of the Development of the Georgia High School Graduation Test

1990–91	• GaDOE issued request for proposals (RFP) for "Services Related to Development of a Test Item Bank to Assess Implementation of Georgia's Quality Core Curriculum (QCC) at the Secondary Level."
1991–92	• Survey for Georgia high school teachers on Quality Core Curriculum objectives (QCC) to be assessed was conducted. • Original blueprints for ELA, Mathematics, Science, Social Studies were generated.
1992–93	• Test specifications for the four content areas were developed. • The initial bank of items was field tested.
1993–94	• First mandatory statewide participation in the GHSGT administration of ELA and Mathematics in spring was administered. • Standard setting for Pass scores in ELA and Mathematics was conducted.
1994–95	• Second year of GHSGT in ELA and Mathematics was administered. Scores counted toward graduation. • Science and Social Studies GHSGT were field tested.
1995–96	• Scores from GHSGT ELA, Mathematics, and Social Studies counted toward graduation. • Standard setting for Pass score in Social Studies was completed. • Science GHSGT was field tested. • GHSGT summer forms in ELA, Mathematics, and Social Studies were prepared.
1996–97	• Operational GHSGT in Science was given in spring administration. • Standard setting for Pass score in Science was completed. • GHSGT fall and winter forms in ELA, Mathematics and Social Studies were developed. • QCC was approved for use in public schools of Georgia.
1997–98	• Standard Setting for Pass Plus scores in ELA, Math, Science and Social Studies was established. • Revised Georgia Quality Core Curriculum (QCC) aligned to GHSGT. • Blueprints were revised.
2003–04	• Enhanced items were added to GHSGT ELA and Mathematics for AYP. • Cut scores for Proficient and Advance Proficient in the enhanced ELA and Mathematics were set and approved.
2005–06	• GHSGT ELA and Science were aligned to the GPS. • Two versions (QCC and GPS) of GHSGT in ELA and Science as transition for students in 2005 spring, 2006 summer and fall administrations were administered and/or prepared.
2006–07	• GHSGT Mathematics and Social Studies will be aligned to the GPS.
2007–08	• ELA and Science GHSGT are completed to align to the GPS.
2008–09	• Social Studies GHSGT aligned to the GPS.
2009–10	• Mathematics GHSGT aligned to the GPS.
2010–11	• Full implementation of GPS in the four content areas will be completed.

Appendix O

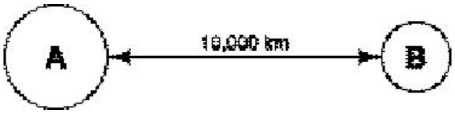
Sample Science Assessment Items Release through Georgia's On-Line Assessment System

Use the table below to answer this question.

	Student 1		Student 2	
	mass	force	mass	force
A	60 kg	30 N	55 kg	35 N
B	80 kg	25 N	70 kg	30 N
C	60 kg	35 N	50 kg	35 N
D	70 kg	30 N	55 kg	25 N

During each of four trials, different students pull on either end of a rope. In which of the circumstances above will the tension in the rope be *greatest*?

(A) A
(B) B
(C) C
(D) D



A student is given the information above about two non-moving objects. Does the student have enough information to calculate the gravitational attraction?

(A) Yes, gravitational force varies with distance.
(B) Yes, since they are not moving, there is no gravitational force.
(C) No, the student does not know their masses.
(D) No, the student does not know what other forces act on them.

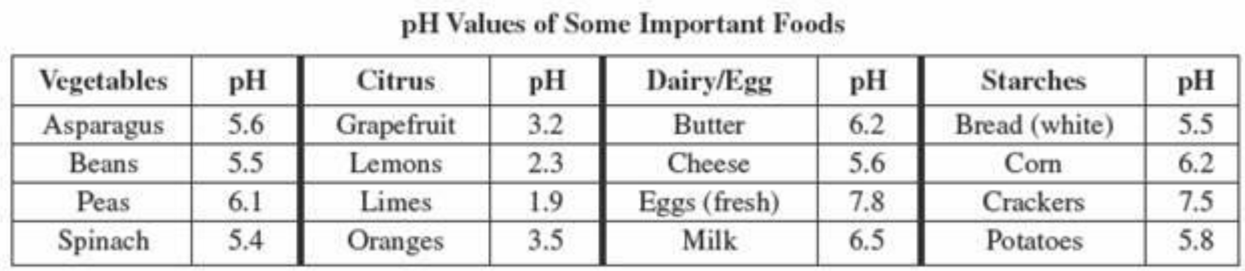
A sound wave is produced and begins to travel from left to right through four different media. The speed of the wave varies as it travels. The media are solid, liquid, gas, and a vacuum, but not necessarily in that order.

1	2	3	4
344 m/sec	5000 m/sec	1450 m/sec	No transmission

Which speed MOST likely represents a gas?

(A) 1
(B) 2
(C) 3
(D) 4

The table below shows pH values of some foods.



pH Values of Some Important Foods

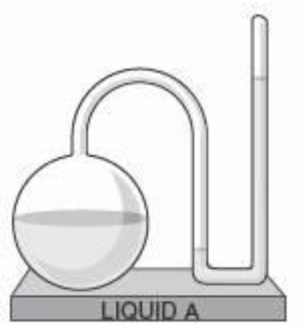
Vegetables	pH	Citrus	pH	Dairy/Egg	pH	Starches	pH
Asparagus	5.6	Grapefruit	3.2	Butter	6.2	Bread (white)	5.5
Beans	5.5	Lemons	2.3	Cheese	5.6	Corn	6.2
Peas	6.1	Limes	1.9	Eggs (fresh)	7.8	Crackers	7.5
Spinach	5.4	Oranges	3.5	Milk	6.5	Potatoes	5.8

A patient has chronic indigestion due to an overproduction of stomach acid. Which foods should the patient avoid until the condition is resolved?

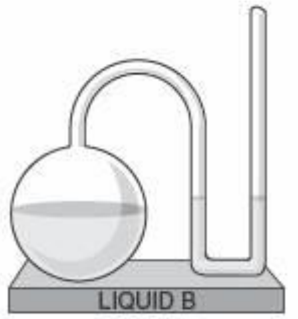
(A) vegetables
(B) citrus
(C) dairy/egg
(D) starches

This online assessment item contains material that has been released to the public by the Massachusetts Department of Education.

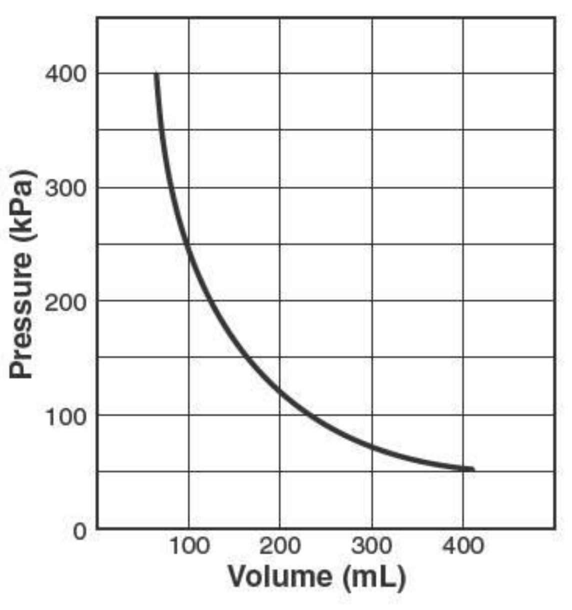
Equal quantities of different liquids are placed in closed manometers at 20°C. Which liquid has the highest vapor pressure?

(A)


(C)

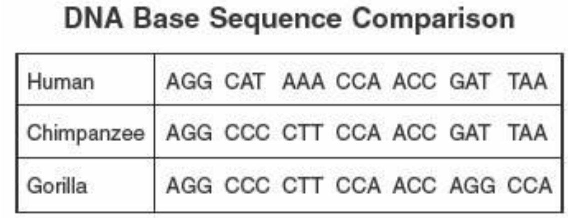

(B)


(D)

The graph shows the pressure of an ideal gas as a function of its volume. According to the graph, increasing the volume from 100 mL to 150 mL —

(A) decreases the pressure by 80 kPa
(B) decreases the pressure by 160 kPa
(C) increases the pressure by 80 kPa
(D) increases the pressure by 160 kPa



DNA Base Sequence Comparison

Human	AGG CAT AAA CCA ACC GAT TAA
Chimpanzee	AGG CCC CTT CCA ACC GAT TAA
Gorilla	AGG CCC CTT CCA ACC AGG CCA

This chart compares the base sequences of homologous segments of DNA from three primates. Based on this information, how many differences in the resulting amino acid sequences would you expect to find between humans and chimpanzees?

(A) 2
(B) 3
(C) 4
(D) 6