2018

Middle Grades Science in Florida: A Comparison of Student Achievement in Comprehensive and Subject-specific Science Courses 2013-2017

Kenneth Moore
University of Central Florida

MIDDLE GRADES SCIENCE IN FLORIDA:
A COMPARISON OF STUDENT ACHIEVEMENT IN COMPREHENSIVE AND SUBJECT-SPECIFIC SCIENCE COURSES 2013–2017

by

KENNETH R. MOORE
B.B.A. Kent State University Ohio, 1988
M.B.A. Kent State University Ohio, 1989
M.M.O.A.S Air Command and Staff College Alabama, 2001

A dissertation submitted in partial fulfillment of the requirements
for the degree of Doctor of Education
in the School of Teaching, Learning, and Leadership
in the College of Education and Human Performance
at the University of Central Florida
Orlando, Florida

Spring Term
2018

Major Professors: Rosemarye T. Taylor
Lee Baldwin

ABSTRACT

As with U.S. student achievement on national and international science assessments, Florida's 8th grade student achievement on the 2013–2017 8th grade Florida Comprehensive Assessment Test (FCAT) 2.0 Science/Statewide Science Assessment (SSA) was stagnant. To break this stagnation, many Florida school districts have changed middle grades science course offerings from traditional, subject-specific, discipline-based, layered, or field-specific science courses to comprehensive, integrated, spiraled, interdisciplinary, multidisciplinary, thematic, or general science courses. There was a lack of research showing if either type of science course improved student achievement on standardized science assessments. Controlling for school district student population, low socio-economic status (SES) student percentage, and English learner (EL) percentage, this study compared the 2013–2017 8th grade FCAT 2.0 Science/SSA school district mean scale scores of two groups of school districts: those that offered comprehensive science courses and those that offered subject-specific science courses. Scores for three student groups were analyzed: all students, low SES students, and ELs. No statistically significant differences were found in school district mean scale scores or pass rates between the two school district groups. The comprehensive group mean scale scores were numerically higher, while the subject-specific group mean pass rates were numerically higher. The subject-specific group had statistically significantly higher raw scores for life science and physical science. The comprehensive group had wider dispersions of mean scale scores and pass rates, suggesting inconsistencies in implementation of comprehensive science courses. The primary implication of this study is that educational leaders should not expect to improve student science achievement simply by changing the type of science course offering. Changes should be made

with consideration to student needs, school district demographics, teacher professional development and support, course structure and coherence with standards, and the need for flexibility in teacher assignments.

I dedicate my dissertation work to my family. To my parents, Ken and Helen Moore; sister, Rhonda; and brother, Perry, I am forever grateful for your love, encouragement, and support through all the years leading to this point. To my sons, Logan and Ian, and grandson, Ezra, your inspiration and understanding have kept me going, and given me purpose. I love all of you more than these words can express.

ACKNOWLEDGMENTS

To my dissertation co-chairs, Drs. Rosemarye Taylor and Lee Baldwin, thank you for your encouragement, advice, and inspiration throughout this process. I am both proud and humbled to have been one of your students. To my committee members, Drs. Sue Gao and Valerie Storey, thank you for your expert guidance and advice over the course of my study. To my Cohort 6 colleagues, I am blessed to have studied among you and to have learned from you. Your passion, dedication, and diversity have helped me grow as an educator and as a human being. I am grateful to the U.S. Air Force and the Veteran's Administration for picking up the tab for this and previous degrees. My blood runs Air Force blue!

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER I: INTRODUCTION

Recent international, national, and state standardized assessments show that science achievement of middle school students in the United States is stagnant (U.S. Department of Education [U.S. ED], 2015). Results of the 2015 Program for International Student Assessment (PISA) for science put the US in the middle of the pack among the 72 participating countries (Kastberg, Chan, & Murray, 2016), and only slightly better than average on the 2015 Trends in International Mathematics and Science Study (TIMSS) (International Association for the Evaluation of Educational Achievement [IEA], 2016). These positions were unchanged from the 2012 PISA and 2011 TIMSS assessments (IEA, 2016; Kastberg et al., 2016). Less than half of 8th grade students achieved the minimum level of proficiency in science on the 2009–2015 National Assessment of Educational Progress (NAEP) and the 2013–2017 8th grade Florida Comprehensive Assessment Test (FCAT) 2.0 Science/Statewide Science Assessment (SSA) (Florida Department of Education [FLDOE], 2012b, 2015e, 2016c; National Center for Education Statistics [NCES], 2017c). The percentages of students that achieved the minimum level of proficiency on these assessments are summarized in Figure 1.

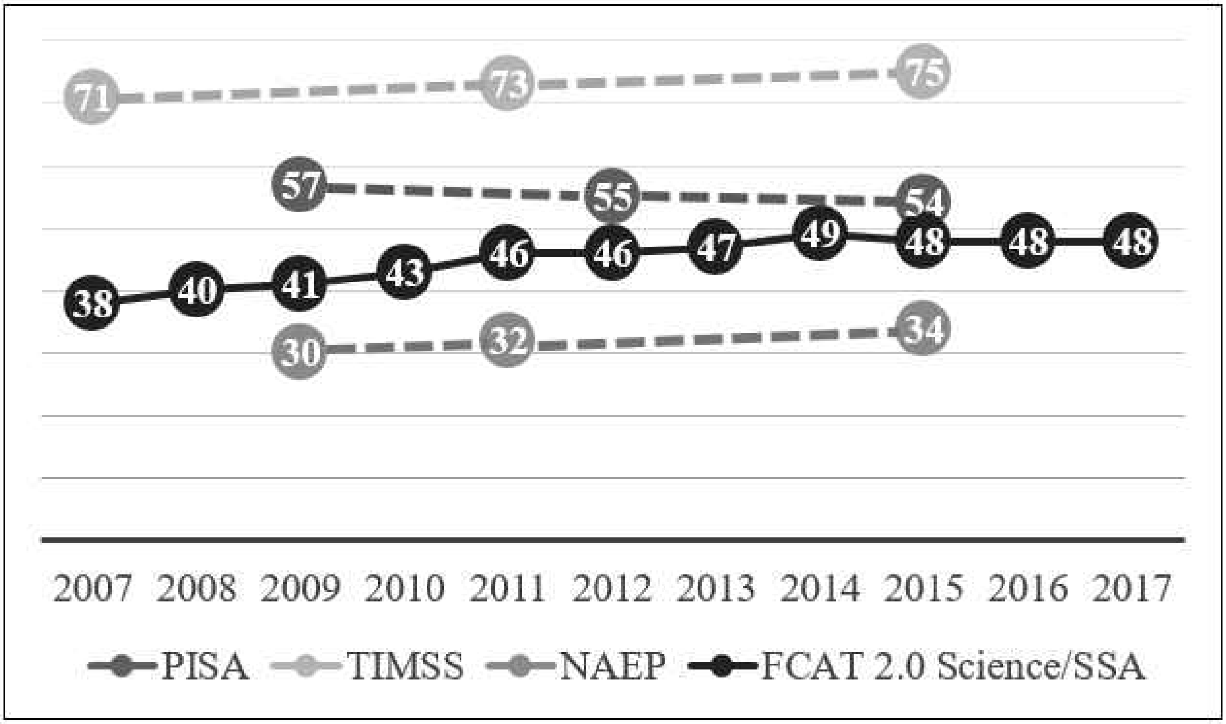

Figure 1. Percent of students, minimum proficiency or higher, 2007–2017.
Dashed lines represent years in which science assessments were not conducted. PISA, TIMSS, and NAEP data compiled from NCES Data Explorer (NCES, 2017e). FCAT 2.0 Science/SSA data for Florida 8th grade students compiled from Florida Standards Assessments Science Results website (FLDOE, 2016c).

This stagnation of science achievement among U.S. students has been deemed a national crisis by some (Langdon, McKittrick, Beede, Khan, & Doms, 2011; National Academy of Sciences [NAS], 2007; National Science Board [NSB], 2007; U.S. ED, 2016b, 2016a). In 2007, the National Science Board (NSB) declared careers and education in science, technology, engineering, and math (STEM) a national priority (NSB, 2007). The U.S. Department of Education projects that by the year 2020, jobs requiring skills in STEM will grow more than 65 percent (U.S. ED, 2016b), while the US has an inadequate number of students becoming proficient in STEM skills to meet the projected demand (Langdon et al., 2011; U.S. ED, 2016b).

The President's Council of Advisors on Science and Technology issued a report to President Obama in 2012 stating that the US must produce a million more STEM professionals by 2022 if the nation is to "retain its historical preeminence in science and technology" (Olson & Riordan, 2012, p. i). Bandura's (1997, 2000) self-efficacy theory suggests that low student achievement on assessments may lead to students' low propensity to pursue STEM careers (Betz, Hackett, Lent, Lopez, & Bieschke, 1991; Hackett & Betz, 1981; National Research Council [NRC], 2012; NSB, 2007; Pajares, 1996).

Other researchers view the STEM career issue as less ominous or even as a manufactured crisis. Michael Anft (2013) states that there is actually significant unemployment among STEM professionals, and that the crisis has been created by technology companies and foundations seeking government funding. Xue and Larson (2015) state that the term STEM is too broad to describe the situation, and that there is no agreement as to whether the term applies only to degreed STEM professionals or more broadly to include lab technicians and skilled trades. A 2014 report by the U.S. Census Bureau, based on 2010 census data, showed that 74 percent of those with a bachelor's degree in STEM were not employed in STEM occupations (U.S. Census Bureau, 2014). Still other researchers state that what is needed is not necessarily an increase in the pipeline of students preparing for STEM careers, but a higher level of science literacy or science awareness for everyone to function as productive, informed citizens (American Association for the Advancement of Science [AAAS], 1989; DeBoer, 2000; Faulkner, 2012; V. J. Mayer, 2002; Shamos, 1995; Wenning, 2007).

Looming crisis or not, national and state education standards and standardized assessments have been revised, new science curricula developed and implemented, and funding for science education research increased in an effort to increase K-12 science achievement (FLDOE, 2017n; NRC, 2012; National Science Foundation [NSF], 2017; M. R. Wilson & Bertenthal, 2006). The No Child Left Behind (NCLB) Act of 2001 required states to develop science education standards and assessments aligned to those standards by 2008. The Every Student Succeeds Act (ESSA) of 2015 (ESSA, 2015) eased some of the accountability measures and penalties of NCLB, but state science and other assessments are still required. The number one goal of the Florida Department of Education is "highest student achievement" (FLDOE, 2014f, para. 2). For school district leaders in Florida, where student achievement on state science assessments is included in school and school district evaluations, improving science achievement is a way to improve evaluation outcomes (FLDOE, 2017a).

The FLDOE prescribes K-12 science education standards known as Next Generation Sunshine State Standards (NGSSS) for science. The NGSSS for science are categorized into 18 general topic clusters related to physical science, life science, Earth and space science, and nature of science (scientific knowledge and practices) (FLDOE, 2017d, 2017n; Florida State University [FSU], 2017). Each topic cluster includes up to four standards, totaling to 68 standards. Each standard includes several benchmarks for grades K–5, 6, 7, 8, and 9–12. There are 109 middle grades science benchmarks, up to 94 of which were assessed on the 2013–2017 8th grade FCAT 2.0 Science/SSA (FSU, 2017).

The FLDOE has approved two general types of middle grades science courses to present these standards: comprehensive and subject-specific (FLDOE, 2017d, 2017n). Each school district decides which type of course to offer and to adopt any curriculum aligned to the standards (§1003.02(1)(d), Fla. Stat., 2016; FLDOE, 2015a). Subject-specific science courses sequence the NGSSS for science by subject area (physical, life, and Earth/space sciences). Nature of science concepts are included in each course. Each subject is offered as a separate course. Earth/space science is typically taught in 6th grade, life science in 7th grade, and physical science in 8th grade, although this sequence is not universal (FLDOE, 2017d). Comprehensive science courses present science standards sequenced not by subject area, but by theme, and are presented each year in grades 6 through 8. Unlike the national Next Generation Science Standards (NGSS), these themes are not specified in Florida's NGSSS for science, and the manner in which the standards are sequenced in comprehensive science courses may vary widely among the school districts (FLDOE, 2017d; Kesidou & Roseman, 2002).

Following recommendations from publications such as *A Framework for K-12 Science Education: Practices, Crosscutting Concepts, and Core Ideas* (NRC, 2012), and the national Next Generation Science Standards (NGSS) (NGSS Lead States, 2013b), the trend in U.S. school districts has been away from traditional, subject-specific science courses to comprehensive, or integrated science courses (Banilower et al., 2013; Czerniak, 2007; FLDOE, 2017d; Hoeg & Bencze, 2017; Huff & Yager, 2016; NRC, 2012). Comprehensive courses are recommended to improve middle grades science achievement because they focus on concepts that span the traditional subject area boundaries, helping students learn the core ideas of each subject area in more depth (NRC, 2012).

In a 2012 National Survey of Science and Mathematics Education, Banilower, Smith, Weiss, Malzahn, Campbell, and Weis (2013) found that most U.S. school districts offered comprehensive science courses only in grade 6, and subject-specific courses only in grades 7 and 8. A percentage of schools offered both as they transitioned from one course type to the other

(Banilower et al., 2013). These data indicate that many schools were transitioning from subject-specific to comprehensive science courses in 2012 (Banilower et al., 2013). Because ESSA requires middle school science assessment in the 8^{th} grade, schools transitioning from one type of course offering to another typically phase in the new courses over three years (Alwardt, 2011; ESSA, 2015; NRC, 2012). Table 1 shows the percentage of U.S. middle schools offering comprehensive and subject-specific science courses as of 2012. An update to this survey is scheduled to be published in late 2018 (Horizon Research, Inc., 2017).

Table 1

2012 U.S. Middle School Science Course Offerings, by Grade (N = 359 schools)

	Grade		
	6	7	8
Science course type	%	%	%
Comprehensive only	45	38	36
Subject-specific only	36	46	47
Both	19	16	17

Note. Adapted from "Report of the 2012 National Survey of Science and Mathematics Education", by E. R. Banilower, et al., 2012. © 2012 by Horizon Research, Inc. Adapted with permission.

The majority of Florida school districts have followed the national trend in offering comprehensive science courses in the middle grades. However, there is no consensus on which type of science course leads to higher student achievement. From the 2012–2013 to the 2016–2017 school years, eleven Florida school districts changed science course offerings. Eight changed from subject-specific to comprehensive courses, and three changed from comprehensive to subject-specific courses (FLDOE, 2013a, 2014b, 2015b, 2016b, 2017f). Four of these changes were made in the 2016–2017 school year. During this time, the number of Florida school districts offering comprehensive middle grades science courses grew from 49 to 50 school districts, and the number of Florida school districts offering subject-specific middle grades science courses fell from 18 school (FLDOE, 2013a, 2014b, 2015b, 2016b, 2017f). Table 2 reflects this data, and Appendix C provides more detail, showing the course offerings of each school district from the 2012–2013 through 2016–2017 school years.

Table 2

Florida School District Middle School Course Offerings (*N* = 67 school districts)

	Florida school districts				
	2012–13[a]	2013–14[b]	2014–15[c]	2015–16[d]	2016–17[e]
Science course type	%	%	%	%	%
Comprehensive	73	78	79	76	76
Subject-specific	27	22	21	24	24

Notes: Data compiled from FLDOE PK–12 Public School Data Publications and Reports. [a]Course Enrollment Survey 3, 2012–13 (FLDOE, 2013). [b]Course Enrollment Survey 3, 2013–14 (FLDOE, 2014b). [c]Course Enrollment Survey 3, 2014–15 (FLDOE, 2015a). [d]Course Enrollment Survey 3, 2015–16 (FLDOE, 2016a). [e]Course Enrollment Survey 3, 2016–17 (FLDOE 2017f).

Scant quantitative research exists to show if either comprehensive or subject-specific science courses improve student achievement (Åström & Karlsson, 2012; Tamassia & Frans, 2014). The literature search, detailed in Chapter 2, yielded only two published, peer-reviewed studies comparing academic achievement of students in comprehensive science courses to students in subject-specific science courses (Åström & Karlsson, 2012; Tamassia & Frans, 2014). Both were conducted in Europe. Three U.S.-based doctoral dissertations on this topic were found (Alwardt, 2011; Clifford, 2016; Faulkner, 2012), as well as a European doctoral dissertation and a master's thesis, also by Åström (2007, 2008).

Six factors about science education in the state of Florida from 2013 to 2017 presented an opportunity to help fill this gap in the research. Science standards remained constant (FLDOE, 2017l). Despite a name change from Florida Comprehensive Assessment Test 2.0 Science (FCAT 2.0 Science) to the Statewide Science Assessment (SSA) in the 2015–16 school year, the assessment remained consistent in content, format, administration, complexity level, scale scoring, and achievement level criteria (FLDOE, 2017l). School district FCAT 2.0 Science/SSA scores were in the public domain (FLDOE, 2017i). School district demographic data were in the public domain. Course surveys showing the science courses offered each year by each school district were in the public domain (FLDOE, 2013a, 2014b, 2015b, 2016b, 2017f). A large sample existed of school districts offering each type of science course among the 67 school districts.

This study compared 8th grade FCAT 2.0 Science/SSA school district mean scale scores for 2013, 2014, 2015, 2016, and 2017 for Florida public school districts that offered comprehensive middle school science courses to those that offered subject-specific middle school science courses for the general education student population. This study also compared the 8th grade FCAT 2.0 Science/SSA school district mean scale scores for low socio-economic status (SES) students and English learners (ELs) in school districts that offered comprehensive middle school science courses to those that offered subject-specific middle school science courses. Not included in this study are scores for Florida's seven special school districts (four

Florida laboratory schools, the Florida School for the Deaf and Blind, Florida Virtual School, and the Oneida Youth Development Center), charter, virtual, and other special schools within Florida's 67 public school districts. While these special, charter, and virtual schools are required to teach to the Florida Next Generation Sunshine State Standards (NGSSS), they are not bound to follow the curricular decisions of the school district in which they operate (§1002.33 (6)(a)2, Fla. Stat., 2016).

Statement of the Problem

The problem studied was the stagnation of science achievement of 8th grade science students as measured by required science assessments in Florida during the middle school years. There was a lack of research on the type of middle school science course, either comprehensive or subject-specific, that resulted in greater school district mean scale scores on the 8th grade FCAT 2.0 Science/SSA in Florida.

Purpose of the Study

The purpose of this study was to determine if there was a difference in student achievement on the 8th grade FCAT 2.0 Science/SSA for 2013, 2014, 2015, 2016, and 2017 between two groups of school districts: those that offered comprehensive middle grades science courses and those that offered subject-specific middle grades science courses. An additional purpose was to ascertain if student demographic characteristics (school district student population, low SES, and EL) were associated with student achievement in school districts that offered either of the two types of science courses.

Significance of the Study

This study is significant in that it helps fill a void in the research by showing if the type of middle grades science course influenced 8th grade FCAT 2.0 Science/SSA school district mean scale scores. This study provides research-based evidence Florida educational leaders may use to make informed decisions about science course offerings and curricula.

Definitions of Terms

Achievement Level, Scale Score, and Pass Rate

Student achievement on the Statewide Science Assessment (SSA, known as Florida Comprehensive Assessment Test 2.0 Science [FCAT 2.0 Science] prior to 2015) is categorized into five achievement levels, with 5 being the maximum, and 3 being the minimum passing level (Fla. Admin. Code R. 6A-1.09422, 2016; FLDOE, 2015a, 2016a). Achievement levels were derived from scale scores ranging from 140 to 260, with 203 being the minimum passing scale score. Although FCAT 2.0 Science/SSA is administered annually, it is not an annual assessment because each student takes it only in grades 5 and 8. Because of this, scale scores are not considered developmental scale scores as in the annual assessments for mathematics and English language arts (Fla. Admin. Code R. 6A-1.09422, 2016; FLDOE, 2015a, 2016a). The pass rate is the percentage of students achieving Level 3 or higher on the FCAT 2.0 Science/SSA (FLDOE, 2017o).

Comprehensive Science Course

A comprehensive science course blends Florida's middle school Next Generation Sunshine State Standards (NGSSS) for nature of science, physical science, life science, and

Earth/space science each year in grades 6 through 8. The focus and sequence of standards is based on scientific practices, concepts, themes, or problems that span the traditional subject-area boundaries. These concepts are revisited throughout each middle grade level in the context of each subject area. According to the NRC (2012), a comprehensive course enables students "over multiple years of school, [to] actively engage in science and engineering practices and apply crosscutting concepts to deepen their understanding of each field's [subject's] disciplinary core ideas" (NRC, 2012, p. 2). Other terms commonly used are integrated, coordinated, spiraled, interdisciplinary, multidisciplinary, thematic, and general science courses (Herr, 2007; NRC, 2012). There are no universally-accepted definitions among the education and scientific communities for these terms (International Bureau of Education [IBE]-UNESCO, 2017b; Ragel, 2015; Stengel, 1997). The term comprehensive science course is used in this study as it is used in Florida's course code directory for middle/junior high (M/J) school science courses, e.g. M/J Comprehensive Science 1, M/J Comprehensive Science 2, and M/J Comprehensive Science 3 (FLDOE, 2017a). For this study, this term refers only to the sequencing and focus of Florida's NGSSS for science in courses offered in grades 6 through 8. This term does not refer to other curriculum components such as instructional methods (inquiry, cooperative learning, direct instruction, etc.), learning experience characteristics (activity/laboratory, textbook, digital, etc.), or specific lesson design (formative/summative assessment, learning goals, differentiated instruction, etc.).

Low Socio-economic Status Student

A low socio-economic status (SES) student is an economically disadvantaged student who is eligible for free or reduced-price meals under the National School Lunch Program

(FLDOE, 2015c). Numbers and percentages of economically disadvantaged students in this study are the school district percentages of students who qualify individually for free or reduced-price lunch, and those students enrolled in a U.S. Department of Agriculture (USDA) Provision 2 or USDA Community Eligibility Provision (CEP) school that serves students from predominantly low-income families (FLDOE, 2015c).

English Learner

The Florida Department of Education (2015c) defines an English learner (EL), or English Language Learner (ELL) as:

> A student who was not born in the US and whose native language is other than English; or was born in the US but who comes from a home in which a language other than English is most relied upon for communication; or is an American Indian or Alaskan Native and comes from a home in which a language other than English has had a significant impact on his or her level of English language proficiency; and who as a result of the above has sufficient difficulty speaking, reading, writing or understanding the English language to deny him or her the opportunity to learn successfully in classrooms in which the language of instruction is English. (p. 2)

Numbers and percentages of EL students in this study include students designated as PK–12 ELL code LY, which means, "the student is an English Language Learner and is enrolled in classes specifically designed for English Language Learners" (FLDOE, 2012a, p.4, 2017c). Students coded LF, LP, LZ, or ZZ are not included (FLDOE, 2012a, 2017c).

Next Generation Science Standards

Next Generation Science Standards (NGSS, not to be confused with Florida's NGSSS for science) are K–12 science standards developed by 26 states, the National Research Council (NRC), the National Science Teachers Association (NSTA), the American Association for the Advancement of Science (AAAS), and other partners (NGSS Lead States, 2013b). These standards, released in 2013, are based on the NRC's *A Framework for K–12 Science Education* (NRC, 2012; NGSS Lead States, 2013). The NGSS include standards for three science subjects: physical science, life science, and Earth/space science, as well as standards for engineering, technology, and the applications of science (NGSS Lead States, 2013b). The standards are organized by grade level for K–5, and grade-banded 6 through 8 for middle and 9 through 12 for high school (NGSS Lead States, 2013b). NGSS are designed for use in comprehensive science courses (NGSS Lead States, 2013b).

Next Generation Sunshine State Standards for Science

Florida's K–12 science standards are known as Next Generation Sunshine State Standards (NGSSS) for science. Adopted in 2008, the middle grades NGSSS for science include standards for nature of science, physical science, life science, and Earth/space science (FLDOE, 2017g). Nature of science standards focus on the overarching processes, practices and concepts of science and scientific knowledge, such as the systematic gathering of information through direct and indirect observation, the testing of data and information by experimentation and investigation methods, the formation and development of scientific knowledge, and the laws and theories related to those concepts (National Science Teachers Association [NSTA], 2005). Florida's NGSSS for science are organized by both grade level and subject area for grades K through 8, and by grade band (9 through 12) and subject area for high school. The same standards are used in both comprehensive and subject-specific middle school science courses (FLDOE, 2017g).

Standards and Benchmarks

Often used interchangeably, these two terms have specific definitions in the state of Florida. Florida's NGSSS for science are categorized into 18 general topic clusters related to physical science, life science, Earth and space science, and nature of science (scientific knowledge, processes, and practices) (FLDOE, 2017d, 2017n; FSU, 2017). Each topic cluster includes up to four standards, totaling to 68 standards. Each standard includes several benchmarks for grades K–5, 6, 7, 8, and 9–12. There are 109 middle grades science benchmarks, up to 94 of which were assessed on the 2013–2017 8th grade FCAT 2.0 Science/SSA (FSU, 2017). The following standard for nature of science is presented as an example.

> Big Idea 1: The Practice of Science
>
> A: Scientific inquiry is a multifaceted activity. The processes of science include the formulation of scientifically investigable questions, construction of investigations into those questions, the collection of appropriate data, the evaluation of the meaning of those data, and the communication of this evaluation.
>
> B: The processes of science frequently do not correspond to the traditional portrayal of "the scientific method."
>
> C: Scientific argumentation is a necessary part of scientific inquiry and plays an important role in the generation and validation of scientific knowledge.

> D: Scientific knowledge is based on observation and inference; it is important to recognize that these are very different things. Not only does science require creativity in its methods and processes, but also in its questions and explanations. (FLDOE, 2012d, p. 20)

A 6th grade benchmark related to this standard is, "SC.6.N.1.3: Explain the difference between an experiment and other types of scientific investigation, and explain the relative benefits and limitations of each" (FLDOE, 2014g, p. 35).

Subject-specific Science Course

A subject-specific science course separates NGSSS for science benchmarks for physical science, life science, and Earth/space science, into separate subjects. One subject is offered in each grade 6 through 8, typically Earth/space science in 6th grade, life science in 7th grade, and physical science in 8th grade. Overarching nature of science standards are included along with each subject in each grade 6 through 8 (FLDOE, 2015a). Other terms commonly used are discipline-based, layered, field-specific, didactic, and traditional science courses (IBE-UNESCO, 2017a). As with the myriad synonyms for comprehensive courses, the distinctions among these terms are not universally accepted (AAAS, 1989; IBE-UNESCO, 2017; Stengel, 1997). The term subject-specific science course is used in this study as the subject is used in Florida's course code directory for middle/junior high (M/J) grades science courses, e.g. M/J Physical Science, M/J Life Science, and M/J Earth/Space Science (FLDOE, 2017d). For this study, this term refers only to the sequencing and focus of Florida's NGSSS for science benchmarks in courses offered in grades 6 through 8. This term does not refer to other curriculum components such as instructional methods (inquiry, cooperative learning, direct instruction, etc.), learning experience characteristics (activity/laboratory, textbook, digital, etc.), or specific lesson design (formative/summative assessment, learning goals, differentiated instruction, etc.).

Conceptual Framework

The need for this study and the selection of the research questions were supported through a literature review focused on five concepts related to middle grades science education and assessment and which provided the conceptual framework for this study. These concepts were science education, science assessment, science standards, science education challenges for large school districts, low socio-economic status (SES) students and English learners (ELs), and related research on comprehensive and subject-specific science courses.

Science Education

First, the status of science education in the US and the state of Florida was reviewed to support the need for this study. Results of international, national, and state standardized assessments, including Program for International Student Assessment (PISA), Trends in International Mathematics and Science Study (TIMSS), National Assessment of Educational Progress (NAEP), and the Florida Comprehensive Assessment Test 2.0 Science (FCAT 2.0 Science)/Statewide Science Assessment (SSA) show that science achievement of middle school students in the US is stagnant (FLDOE, 2017k; IEA, 2016; Kastberg et al., 2016; NCES, 2017b). The implications of these scores were explored from national, student, and Florida educational leadership perspectives.

Science Assessment

Standardized assessment of student achievement on middle grades science standards was the second concept supporting this study. The frameworks and science standards assessed in the PISA, TIMSS, NAEP, and FCAT 2.0 Science/SSA science assessments are compared. This section concludes with a look at how standardized assessments may affect student performance.

Science Standards

The third concept underpinning this study was related to science standards at the international, national, and state levels. First, the historical background of the development of science standards was explored. Following the historical background, documentation on the consistency and alignment of Florida's NGSSS for science to FCAT 2.0 Science/SSA was reviewed.

Science Education Challenges

School districts face many challenges when implementing new science curriculum and course offerings. These challenges comprised the fourth concept underpinning this study, and was the basis for Research Questions 2, 3, and 4. Larger school districts often experience inconsistency in the implementation of new course offerings due to varying attitudes, beliefs, education, professional development, and subject area certifications among the teaching force (Davis, 2003; Diehl, 2005). Student factors such as poverty and lack of English language skill have been shown to influence student performance on standardized assessments, presenting special challenges for low socio-economic status (SES) students and English learners (ELs) (Amaral, Garrison, & Klentschy, 2002; Driscoll, Halcoussis, & Svorny, 2003; Lippman, Burns,

& McArthur, 1996; Maerten-Rivera, Ahn, Lanier, Diaz, & Lee, 2016; Matkins, McDonnough, & Henschel, 2014; Wiseman, 2012).

Science Courses

The fifth and final concept of the framework was an analysis of science courses that support student learning. Extensive searches yielded only two published, peer-reviewed, academic studies that compared comprehensive to subject-specific science courses (Åström & Karlsson, 2012; Tamassia & Frans, 2014). These relevant empirical studies were reviewed, as well as several doctoral dissertations comparing achievement of students in comprehensive and subject-specific science courses.

Research Questions

There was a lack of definitive research to show if either type of science course (comprehensive or subject-specific) led to higher 8th grade FCAT 2.0 Science/SSA school district mean scale scores. Four Research Questions were chosen to fill this gap in the research. All Research Questions used the type of school district science course offering (comprehensive or subject-specific) as the independent variable, and the 8th grade FCAT 2.0 Science/SSA school district mean scale scores as the dependent variable.

1. To what extent did 8th grade FCAT 2.0 Science/SSA school district mean scale scores for 2013, 2014, 2015, 2016, and 2017 differ between two groups of Florida school districts: those that offered comprehensive science courses and those that offered subject-specific science courses?

2. To what extent did 8th grade FCAT 2.0 Science/SSA school district mean scale scores for 2013, 2014, 2015, 2016, and 2017 differ between two groups of Florida school districts: those that offered comprehensive science courses and those that offered subject-specific science courses, when the school districts are matched by overall population size, low SES student percentage, and EL percentage?

3. To what extent did 8th grade FCAT 2.0 Science/SSA school district mean scale scores for low SES students for 2013, 2014, 2015, 2016, and 2017 differ between two groups of Florida school districts: those that offered comprehensive science courses and those that offered subject-specific science courses, when the school districts are matched by overall population size, low SES student percentage, and EL percentage?

4. To what extent did 8th grade FCAT 2.0 Science/SSA school district mean scale scores for ELs for 2013, 2014, 2015, 2016, and 2017 differ between two groups of Florida school districts: those that offered comprehensive science courses and those that offered subject-specific science courses, when the school districts are matched by overall population size, low SES student percentage, and EL percentage?

Research Question 1 analyzed 8th grade FCAT 2.0 Science/SSA school district mean scale score differences using all Florida school district data for all 8th grade students assessed, low SES students, and ELs, without controlling for demographic that may affect student achievement, such as school district student population, low SES, or EL status. Research Questions 2, 3, and 4 controlled for these demographic factors. The 8th grade FCAT 2.0 Science/SSA school district mean scale scores differences were analyzed in paired samples of school districts that offered different types of science courses, but similar student populations, percentages of low SES students, and percentages of ELs. Research Question 2 analyzed the differences in school district mean scale scores for of all students in the paired school district samples. Research Questions 3 analyzed the differences in school district mean scale scores for low SES students in the paired school district samples. Research Question 4 analyzed the differences in school district mean scale score for ELs in the paired school district samples.

Research supports the selection of school district student population, percentage of low SES students, and percentage of ELs as control factors for this study. Lippman, Burns, McArthur and the NCES (1996) and Driscoll, Halcoussis, and Svorny (2003), and McLaughlin (2014) have documented a negative effect of large school and school district size on student performance on standardized assessments, even when controlling for socio-economic factors. Other research has shown that special teacher preparation and instructional intervention is needed in large, urban school districts (Matkins et al., 2014; Schindel Dimick, 2016; White, Brown, Viator, Byrne, & Ricchezza, 2017).

In school districts of all sizes, the negative effects of poverty on student achievement are well documented. The NAEP 8th grade science assessment scores have shown a consistently wide achievement gap between economically disadvantaged students and non-economically disadvantaged students since 2009 (NCES, 2015a). On the 2016 SSA, only 39% of economically disadvantaged students achieved the minimum level of proficiency on Florida's 2016 SSA, as contrasted with the 65.9% for non-economically disadvantaged students (FLDOE, 2017e). Wiseman (2012), and Miller, Votruba-Drzal, and Seodji (2013) concluded that student poverty is a strong predictor of low science achievement in countries around the world.

Research Question 3 was chosen to assess whether either type of science course led to higher 8th grade FCAT 2.0 Science/SSA school district mean scale scores for low SES students.

As with low SES students, the achievement gap between EL and non-EL students is well documented (González-Howard & McNeill, 2016; Houseal, Gillis, Helmsing, & Hutchison, 2016; MacDonald, 2004). Special intervention programs have helped narrow this gap (Amaral et al., 2002; Maerten-Rivera et al., 2016), but in Florida, the EL achievement gap remains wide. In 2016, only 14.6% of EL students achieved the minimum level of proficiency on the 8th grade SSA, as contrasted with 52.9% of non-EL students (FLDOE, 2017e). Visone (2009, 2010) found that much of the achievement gap for low SES students and ELs can be attributed to the readability of the academic language in the assessments for students with low reading skill levels. Research Question 4 was chosen to assess whether either type of science course led to higher school district mean scale scores on the 8th grade FCAT 2.0 Science/SSA for ELs.

Methodology

This quantitative study compared 8th grade Florida Comprehensive Assessment Test (FCAT) 2.0 Science and Statewide Science Assessment (SSA) school district mean scale scores for 2013, 2014, 2015, 2016, and 2017 of Florida school districts that offered comprehensive science courses to those that offered subject-specific science courses for the middle grades. The range of assessment years used in this study was chosen because, during those years, Florida's Next Generation Sunshine State Standards (NGSSS) for science, the 8th grade FCAT 2.0 Science/SSA, and the scale on which students were scored, remained consistent. An additional factor in the selection of this range of assessment years was the availability of Course Enrollment

Surveys, which enabled the determination of each school district's science curriculum type (comprehensive or subject-specific).

The independent variable for all Research Questions was the school district science course type: comprehensive or subject-specific. The dependent variable for all Research Questions was the 8th grade FCAT 2.0 Science/SSA school district mean scale score from 2013 through 2017. Research Question 1 examined the school district mean scale scores for all students, low SES students, and ELs, in all Florida school districts, using independent school district samples without controls. Research Questions 2, 3, and 4 used paired-samples of school districts, with controls for school district student population, percentage of low SES students, and percentage of ELs. Each school district pair consisted of one school district that offered comprehensive science courses and one that offered subject-specific science courses, matched by school district student population, percentage of low SES students, and percentage of ELs. Research Question 2 analyzed the school district mean scale scores for all students in the paired samples. Research Question 3 examined the school district mean scale scores for only low SES students in the paired samples. Research Question 4 examined the school district mean scale scores for only ELs in the paired samples.

Target Population

The target population was Florida's 8th grade general education students from traditional public (non-charter, non-virtual, and non-special) middle and junior high schools who took the FCAT 2.0 Science/SSA each year from 2013 through 2017, approximately 160,000 students per year. Students in Florida's four laboratory school districts and three special school districts were excluded from the target population. Likewise, students in charter, virtual, and special schools (correctional facilities, behavioral centers, hospital/homebound, etc.) within the 67 Florida public school districts were excluded from the target population. Two population subgroups were also targeted: low SES students (approximately 95,000 students per year) and ELs (approximately 9,800 students per year).

Data Anonymity

The University of Central Florida's Institutional Review Board determined this study does not constitute human research (see Appendix A). Only public record, school-district-level data were used in this study. These data were obtained from Florida's PK-12 Education Information Portal (FLDOE, 2017i). To prevent disclosure of personal data and protect individual privacy, Florida's PK-12 Education Information Portal contains no individual student identifiers, and suppresses data for schools and school districts with fewer than ten students assessed in any category (FLDOE, 2017i). Since this is not a study of school district results, but a study of approaches to middle school science curriculum, fictitious school district names have been used in this study (Fraenkel, Wallen, & Hyun, 2015). Fictitious names were obtained from an online random name generator (Behind the Name, 2017), and randomly assigned to each school district.

Sampling Method

The sampling method was non-probability, purposive, and stratified (Fraenkel et al., 2015). Because there were unequal numbers of school districts offering either type of science course each year, and because the demographics of the school districts varied, random sampling was not appropriate for finding qualified samples for this study (Fraenkel et al., 2015). The sampling unit and unit of analysis were the school district. The school district was chosen as the sampling unit and unit of analysis for four reasons. First, student-level data were not publicly available (FLDOE, 2017i). Second, school-level data were unavailable for some smaller schools with small numbers of low SES students and ELs. Data are suppressed in the FLDOE EdStats system for any score category comprised of ten or fewer students (FLDOE, 2017i). Third, the type of science course used by schools is determined at the school district level (FLDOE, 2008a). Fourth, the school district unit of analysis facilitated the pairing of school districts that offered comprehensive science courses to school districts that offered subject-specific science courses, matched by school district student population, percentage of low SES students, and percentage of ELs.

For each year 2013 through 2017, publicly-available, archival 8th grade FCAT 2.0 Science/SSA score and demographic data were retrieved from the Florida Department of Education (FLDOE) website (FLDOE, 2017p, 2017i). These data included school district mean scale scores for all students and student subgroups, student population, percentage of low SES students, and percentage of ELs, by school, for each school in each Florida school district. To ensure only scores for general education students in traditional public schools were included, data were removed for schools in Florida's four laboratory school districts and three special school districts. From the remaining data for schools in Florida's 67 public school districts, data were removed for charter, virtual, and other specialized schools within each school district. The remaining school-level mean scale scores were averaged to obtain school district mean scale scores.

The science curriculum type used by each school district for each of these years was determined using publicly-available, archival FLDOE course enrollment surveys (FLDOE, 2013a, 2014b, 2015b, 2016b, 2017f). School districts were divided into two groups, based on the type of science course offered. In this study, these school district groups are referred to as comprehensive groups and subject-specific groups.

Research Question 1 used school district data for all 67 Florida school districts, by year, as samples to analyze differences in 8^{th} grade FCAT 2.0 Science/SSA school district mean scale scores for all students, low SES students, and ELs, between the two school district groups. Research Questions 2, 3, and 4 used paired samples of Florida school districts for each year. These paired samples consisted of school districts that offered different science courses (comprehensive or subject-specific), matched by overall student population, low SES student percentage, and EL percentage for each year from 2013 through 2017. Because the number of school districts that offered subject-specific science courses was smaller than the number of school districts that offered comprehensive science courses, all subject-specific school districts were selected for the subject-specific group in each year's paired sample. The comprehensive group school districts were selected by matching each subject-specific school district to a comprehensive school district with similar student population, percentage of low SES students, and percentage of ELs. Pearson *r* correlation tests were conducted for each demographic matching factor to verify that the paired samples were matched at a statistically significant (α = .01) level (Steinberg, 2011). Research Question 2 analyzed the differences in 8^{th} grade FCAT 2.0 Science/SSA school district mean scale scores for all students in the paired samples. Research Question 3 analyzed 8^{th} grade FCAT 2.0 Science/SSA school district mean scale scores for only low SES students in the paired samples. Research Question 4 analyzed the 8th grade FCAT 2.0 Science/SSA school district mean scale scores for only ELs in the paired samples.

Sample Sizes

Research Question 1 compared, by year, the 8th grade FCAT 2.0 Science/SSA school district mean scale for three groups of students (all students, low SES students, and ELs), of all comprehensive school districts to that of all subject-specific school districts. The school district groups in each year's independent sample were unequal because fewer school districts offered subject-specific science courses. The Research Question 1 school district sample sizes for each year (*N*) and sizes of sample school district groups (*n*) are shown in Table 3, along with the numbers of students assessed represented in the school district samples.

Table 3

RQ 1 School District Sample Sizes by Year

Year	School district group	*n*	Students assessed
2013	Comprehensive	49	108,012
	Subject-specific	18	56,999
		N = 67	165,011
2014	Comprehensive	52	109,891
	Subject-specific	15	56,309
		N = 67	166,200
2015	Comprehensive	53	115,058
	Subject-specific	14	51,552
		N = 67	166,610
2016	Comprehensive	52	107,991
	Subject-specific	15	49,396
		N = 67	157,387
2017	Comprehensive	49*	102,729
	Subject-specific	17	52,237
		N = 66	154,966

Note. *One school district reported no 8th grade FCAT 2.0 Science/SSA results for 2017.

Because fewer school districts offered subject-specific than comprehensive science courses, the paired sample sizes for Research Questions 2, 3, and 4 were limited to the number of school districts that offered subject-specific courses each school year. All school districts that offered subject-specific courses were selected for each year's subject-specific sample group. Each of these was paired with a comprehensive school district, matched by school district student population, percentage of low SES students, and percentage of ELs. The school district paired sample sizes, by year, for Research Questions 2 and 3 are shown in Table 4, along with the total numbers of students assessed and low SES students assessed represented in the samples.

Table 4

RQ 2 and 3 School District Paired Sample Sizes, by Year

Year	*N* (school district pairs)	Total students assessed	Low SES students assessed
2013	18	111,320	64,056
2014	15	111,804	63,220
2015	14	109,183	62,967
2016	15	101,349	59,590
2017	17	108,273	63,018

Research Question 4 analyzed the school district mean scale scores for ELs in the school district in the paired samples. Due to data suppression of mean scale scores for small school districts with fewer than 10 ELs assessed (FLDOE, 2017e), not all the school district paired samples could be used, further limiting the sample sizes. The school district paired sample sizes, by year, for Research Question 4.are shown in Table 5, along with the numbers of ELs assessed represented in the samples.

Table 5

RQ 4 School District Paired Sample Sizes, by Year

Year	*N* (school district pairs)	ELs assessed
2013	9	5,223
2014	9	5,677
2015	9	6,091
2016	9	6,401
2017	11	7,471

Table 6 shows the 2017 Research Questions 2 and 3 school district paired samples (using fictitious school district names) and the demographics used to match each pair. The Research Question 4 paired samples were subgroups of the Research Question 2 and 3 paired samples. School districts with fewer than 10 ELs assessed were excluded from the Research Question 4 samples due to data suppression of the EL mean scale scores (FLDOE, 2017e). School district paired samples for 2013 through 2016 for all Research Questions are shown in Appendix D.

Table 6

RQ 2 and 3 School District Paired Samples, 2017 (*N* = 17 school district pairs)

Pair	School district[a]	Course type[b, c]	Student population[d]	Low SES (%)[e]	EL (%)[f]
1	Phokas	Comprehensive	11,542	48.3	1.1
1	Lucas	Subject-specific	5,010	45.6	0.3
2	Firdaus	Comprehensive	5,266	58.7	9.2
2	Viktor	Subject-specific	8,582	49.4	10.5
3	Roel	Comprehensive	9,173	52.7	3.7
3	Cornell	Subject-specific	10,067	57.2	1.3
4	Amias	Comprehensive	2,752	55.3	2.1
4	Lawson	Subject-specific	8,601	58.9	3.0
5	Desta	Comprehensive	5,500	60.3	4.5
5	Katlyn	Subject-specific	6,056	59.9	5.6
6	Blythe	Comprehensive	4,906	64.9	9.1
6	Emmett	Subject-specific	1,268	57.3	7.1
7	Kimberly	Comprehensive	37,052	44.4	2.2
7	Linwood	Subject-specific	29,485	48.5	2.6
8	Lavender	Comprehensive	28,027	47.4	3.1
8	Renato	Subject-specific	31,091	45.2	3.4
9	Gottfried	Comprehensive	12,930	55.1	2.9
9	Samson	Subject-specific	15,925	50.6	2.5
10	Ross	Comprehensive	73,446	50.6	3.6
10	Sulayman	Subject-specific	67,816	47.2	5.0
11	Ciara	Comprehensive	72,490	54.8	4.3
11	Adil	Subject-specific	42,801	47.8	6.7
12	Lee	Comprehensive	92,682	51.9	9.8
12	Barnaby	Subject-specific	46,407	59.6	15.2
13	Everett	Comprehensive	48,892	54.5	12.9
13	Mirela	Subject-specific	63,023	57.0	19.6
14	Katharine	Comprehensive	40,417	65.2	8.9
14	Placido	Subject-specific	42,516	60.9	4.8
15	Junayd	Comprehensive	63,100	64.4	6.5
15	Walker	Subject-specific	43,040	65.1	6.5
16	Husniya	Comprehensive	214,402	58.6	12.7
16	Giselle	Subject-specific	271,828	61.9	12.7
17	Sunita	Comprehensive	192,729	59.2	12.7
17	Rukiye	Subject-specific	200,667	65.6	14.4

Notes: [a]Fictitious school district names. [b]Course Enrollment by School, Survey 3, 2015–16 (FLDOE, 2017f). [c]Comprehensive = 1, Subject-specific = 2. [d]Student Enrollment by District 2012–17 (FLDOE, 2017e). [e]Economic Status by District 2012–17 (FLDOE, 2017e). [f]ELL Students by District 2012–17 (FLDOE, 2017e).

Test Methods

Research Question 1 used independent samples *t*-tests to determine the statistical significance of the difference in school district 8th grade FCAT 2.0 Science/SSA mean scale scores for all students, low SES students, and ELs, between the comprehensive and subject-specific school district groups, for each year from 2013 to 2017, without regard to school district population, percentage of low SES students, or percentage of ELs. Sample groups of school districts that offered each type of science course for were unequal because about two-thirds of school districts offered comprehensive science courses while only about one-third offered subject-specific science courses from 2013 to 2017. Data were analyzed for normality of distribution (skewness, kurtosis, and outliers) and equality of variances (using Levene's test) to support the validity of the independent samples *t*-tests (Steinberg, 2011). Independent samples *t*-tests are robust to unequal sample sizes and unequal variance in the samples (Kohr & Games, 1974; Steinberg, 2011). When Levene's test did not confirm the equality of variance in the samples, the degrees of freedom used in the independent samples *t*-tests were reduced to account for the inequality (Kohr & Games, 1974; Steinberg, 2011). Also, when only two groups are being analyzed, the Statistical Package for the Social Sciences (SPSS®) independent samples *t*-test for unequal variances provides the same results as Welch's *t*-test (IBM Support, 2016), which is very robust to both unequal sample sizes and variances (Kohr & Games, 1974). For statistically significant differences, effect sizes were calculated using Cohen's *d* (Steinberg, 2011).

While robust to inequalities of variance and sample sizes, independent samples *t*-tests are do not control for other factors that may impact the distributions being tested (Steinberg, 2011).

For this reason, Research Questions 2, 3, and 4 used paired samples *t*-tests to control for three demographic factors that have been shown to impact student achievement: school district student population (Amah, Daminabo-Weje, & Dosunmu, 2013; Driscoll et al., 2003; Lippman et al., 1996), low socio-economic status (SES) (Becker & Luthar, 2002; Hanushek, 2010; Hattie, 2009; Ladd, 2012; Miller et al., 2013; Wiseman, 2012), English learner (EL) status (Cosentino de Cohen, Deterding, & Clewell, 2005; NCES, 2016). Paired samples *t*-tests (Steinberg, 2011) were used to analyze 8th grade FCAT 2.0 Science/SSA school district mean scale score differences in school districts that offered comprehensive science courses and those that offered subject-specific science courses. By year, each subject-specific school district was paired with a comprehensive school district, matched by school district student population, percentage of low SES students, and percentage of ELs. Pearson *r* correlations were calculated to verify that the pairs were matched at a statistically significant level by each demographic factor. Research Question 2 used paired samples *t*-tests (Steinberg, 2011) to analyze differences in school district mean scale scores for all students in the paired school district samples. Research Question 3 used paired samples *t*-tests (Steinberg, 2011) to analyze differences in school district mean scale scores for low SES students in the paired school district samples. Research Question 4 used paired samples *t*-tests (Steinberg, 2011) to analyze differences in school district mean scale scores for ELs in the paired school district samples. All school district mean scale score data were analyzed for normality of distribution (skewness, kurtosis, and outliers) and homogeneity of variance (using Pittman-Morgan tests) to support the validity of the paired samples *t*-tests variance (Gardner, 2001; Kohr & Games, 1974; Morgan, 1939; Pitman, 1939; Steinberg, 2011).

Limitations

1. The results may not be generalizable to other states that offer different types of science courses, prescribe significantly different science standards and benchmarks for the middle school grades, or employ different assessment schedules.

2. The results of this study may not be generalizable to charter or private schools.

3. School district sample sizes (up to $N = 67$) were small relative to the number of students in the target population (approximately 160,000 students per year) (Krejcie & Morgan, 1970). While all 67 Florida school districts were used as the samples for Research Question 1, Research Questions 2, 3, and 4 used school district paired samples, which further limited the sample sizes. Because the number of school districts that offered subject-specific science courses was smaller than the number of school districts that offered comprehensive science courses each year, the paired sample sizes were limited to the number of school districts that offered subject-specific science courses.

4. Overlap exists among the three demographic groups of students (all assessed students, low SES students, and ELs) that comprised the school district mean scale scores analyzed in this study. The school district mean scale scores for all assessed students includes the scores of low SES students, ELs, and all other demographic groups. The school district mean scale scores for low SES students includes the scores of all students who qualify individually for free or reduced-price lunch, and those students enrolled in a U.S. Department of Agriculture (USDA) Provision 2 or USDA Community Eligibility Provision (CEP) school that serves students from predominantly low-income families (FLDOE, 2015c), regardless of other demographic groups to which they may belong. Likewise, the school district mean scale scores for ELs includes all

students designated as PK–12 ELL code LY (FLDOE, 2012a, 2017c), regardless of other demographic groups to which they may belong.

Delimitations

1. This study is focused only on 8th grade FCAT 2.0 Science/SSA scale scores for general education students in the 67 Florida public school districts, not including the special school districts for the four Florida laboratory schools, the Florida School for the Deaf and Blind, Florida Virtual School, and the Oneida Youth Development Center.

2. This study does not include FCAT 2.0 Science/SSA scores for charter schools.

3. This study does not address other factors that may impact student achievement on FCAT 2.0 Science/SSA, such as specific curriculum content, teacher experience; professional development, instructional methods, instructional quality, instructional materials, etc.

4. Scale scores for the four Florida laboratory schools, the Florida School for the Deaf and Blind, Florida Virtual School, the Oneida Youth Development Center, charter schools, virtual schools, and other special schools within the 67 school districts, such as correctional facilities, behavioral centers, exceptional student education centers, hospital/homebound schools, etc., are not included in this study. While these schools are required to teach to the Florida NGSSS for science, they are not bound to follow the curricular decisions of the school district in which they operate (§1002.33 (6)(a)2, Fla. Stat., 2016).

5. The school district was chosen as the sampling unit due to availability of public domain data. All 67 Florida school districts were used as the samples for Research Question 1. Research Questions 2, 3, and 4 used school district paired samples. Each pair consisted of a school district that offered comprehensive science courses and a school district that offered

subject-specific science courses, matched by student population, percentage of low SES students, and percentage of ELs. The matching controlled for demographic factors that may impact student achievement and ensured comparison of school districts that differed only in the type of science courses offered. Because the number of school districts that offered subject-specific science courses was smaller than the number of school districts that offered comprehensive science courses each year, the paired sample sizes were limited to the number of school districts that offered subject-specific science courses.

6. While achievement gaps among three groups of students (all assessed students, low SES students, and ELs) are evident in the analyses presented in this study, this study does not address achievement gaps per se.

7. Of Florida's 67 public school districts, 52 offered elective middle grades science research courses in addition to their normal middle grades science courses, whether comprehensive or subject-specific. The purpose of these courses was to enable students to develop knowledge and skills in scientific research, with emphasis on determining and refining research questions; research design; and data collection and analysis (FSU, 2017). These courses were offered only in a limited number of middle/junior high schools within these school districts, and student enrollment ranged from less than one percent to just over ten percent of school district total student population in the 2016–2017 school year (FLDOE, 2017f). Differences in student achievement between those enrolled and not enrolled in science research courses were not considered in this study.

8. This study is delimited to consideration of only middle school science courses that address Florida's NGSSS for science for grades 6 through 8. Data were not collected for middle

schools that offered high school science courses to students as part of Florida's middle school acceleration program. This program, which is part of Florida's school grading system, allows middle schools to offer a high school biology course in lieu of the normal middle grades comprehensive or subject-specific science courses (FLDOE, 2017a, 2017m). The high school biology course includes a state standardized end-of-course (EOC) assessment. Students' EOC scores may earn a middle school up to an additional 100 points in its school grade calculation (FLDOE, 2017a, 2017m).

Assumptions

1. Each Florida school district adheres to Florida's NGSSS for science whether it offers subject-specific or comprehensive science courses.

2. Because students are assessed only in 8^{th} grade for NGSSS for science taught in grades 6 through 8, it was assumed that any change of course offering implemented by a school district was phased in over three years to avoid any gaps in the NGSSS for science taught to the 8^{th} grade students assessed on the FCAT 2.0 Science/SSA.

3. Because Florida law requires all traditional public schools to test at least 95 percent of eligible students (§1008.34 Fla. Stat., 2016; Fla. Admin. Code R. 6A-1.09422, 2016), the available FCAT 2.0 Science/SSA data analyzed were assumed to represent at least 95 percent of the target population of this study.

4. The independent samples *t*-tests and paired samples *t*-tests used in this study assume normality of distribution and equality of variance in the compared samples. For both the independent samples and paired samples *t*-tests, skewness and kurtosis of the samples were analyzed to ensure normality of the distribution, i.e., skew $< |2.0|$ and kurtosis $< |9.0|$ (Schmider,

Ziegler, Danay, Beyer, & Bühner, 2010). Levene's test (George & Mallery, 2010) was used to test the equality of variances in the samples. For the paired samples *t*-tests, Pitman-Morgan tests (Gardner, 2001; Morgan, 1939; Pitman, 1939) were used to test the assumption of homogeneity of variance in the paired samples.

Organization of the Study

This study is organized in five chapters. Chapter I includes the background of the study, statement of the problem, purpose of the study, significance of the study, definition of terms, conceptual framework, Research Questions, limitations, delimitations, and the assumptions of the study. Chapter II presents a review of the literature. The review includes research on the status of science education; science assessment; science standards; science challenges for large school districts, low SES students, and ELs; and comprehensive and subject-specific science courses. Chapter III describes the methodology used for this research study, procedures used in the selection of samples, and the statistical analysis procedures used to analyze the data for each Research Question. Chapter IV presents the study's findings including the results of statistical analyses, tests of research hypotheses, and confirmatory factor analysis. Chapter V provides a summary of the entire study, discussion of the findings, implications of the findings for theory and practice, recommendations for further research, and conclusions.

Summary

The purpose of this quantitative study was to determine if there was a difference in student achievement on FCAT 2.0 Science/SSA between two groups of Florida school districts: those that offered comprehensive science courses and those that offered subject-specific science courses, from 2013 through 2017. Data were analyzed to ascertain overall results and results for school districts matched by overall student population, low SES student population, and EL student population. The results of this study offer Florida school districts research-based information helpful in making informed decisions about middle school science course offerings.

CHAPTER II: LITERATURE REVIEW

Introduction

The purpose of this study was to determine if there was a difference in student achievement on Florida's 8th grade Florida Comprehensive Assessment Test 2.0 for Science (FCAT 2.0 Science)/Statewide Science Assessment (SSA) between two groups of Florida school districts: those that offered comprehensive science courses and those that offered subject-specific science courses, from 2013 through 2017. The results of this study may provide educational leaders with evidence on which to base decisions regarding middle school science course offerings. The primary research question focused on school district mean scale score differences between groups of school districts that offered either type of science course. The remaining research questions focused on school district mean scale score differences between pairs of school districts that offered different types of science courses, matched by overall student population, percentage of low socio-economic status (SES) students, and percentage of English learner (ELs).

This literature review synthesizes research on five concepts related to the research questions. The first section, science education, presents the status of science education based on the results of the Program for International Student Assessment (PISA), Trends in International Mathematics and Science Study (TIMSS), National Assessment of Educational Progress (NAEP), and FCAT 2.0 Science/SSA. This section includes research on the implications of the results of these assessments from the perspectives of the U.S. science, technology, engineering, and mathematics (STEM) workforce pipeline, scientific literacy, and Florida educational leadership. The second section, science assessment, outlines and compares the frameworks, standards, and cognitive demands of the PISA, TIMSS, NAEP, and FCAT 2.0 Science/SSA, and reviews literature on the impact of standardized assessment from a student perspective. The third section, science standards, reviews research on the influence of science standards on scientific literacy. The fourth section, science challenges, presents research on the challenges of science education for large school districts, low SES students, and ELs.

The concluding section, science courses, examines related research comparing science achievement of students in comprehensive and subject-specific science courses. The literature search yielded only two published, peer-reviewed, academic studies, four doctoral dissertations, and one master's degree thesis. At the end of each section is a table of the topics and citations used in the section. Table 9 shows the search terms and databases used for this literature review.

Table 7

Search Terms and Databases

Databases	Search terms
EBSCOhost (Academic Search Premier, ERIC Education Source, MAS Ultra-School, ProQuest, ProQuest Middle Search Plus, ProQuest Primary Search, ProQuest Education Journals); Google Scholar; JSTOR; Omnifile; Professional Development Collection; PsycINFO; ScienceDirect; Scopus; Teacher Reference Center; UCF Libraries Catalog; Web of Science (Thomson Reuters); Wiley Online Library	Assessment; Cognitive demand; Comprehensive science; Curriculum development; Disciplinary science; Discipline-focused science; Diversity (student); Economically disadvantaged students; Educational leadership; Educational administration; Educational assessment; English learners; FCAT 2.0 science; Field-focused science; Field-specific science; Florida Comprehensive Assessment Test for science; Florida science assessment; Florida statewide science; General science; Grades 6-8/middle grades/middle school/junior high school science; Integrated curriculum; Integrated science; Interdisciplinary approach in education; Interdisciplinary science; Large school districts; Middle/junior high, multidisciplinary science; NAEP; Parental expectations; Performance based assessment; PISA; Professional development; Reciprocal effects model; School district consolidation; Science curriculum; Science education; Science education; Science education standards; Science instruction; Science learning progression; Science self-efficacy; Scientific literacy; Scientific/science literacy; Spiral science curriculum; Standardized science assessment; Standardized tests; Statewide Science Assessment; STEM attrition; STEM pipeline; STEM shortage; STEM workforce; Student evaluation; Subject-specific science; Teacher attitudes; Teaching methods; Teaching practices; Testing problems; Thematic science; Theme-based science; TIMSS; Traditional science

Science Education

This section reviews the status of U.S. and Florida science education and the need for improvement by looking at results of four standardized assessments. The assessments considered are the science achievement sections of the Program for International Student Assessment (PISA), the Trends in International Mathematics and Science Study (TIMSS), the National Assessment of Educational Progress (NAEP), and Florida's standardized science assessment (FCAT 2.0 Science/SSA). The implications of the results of these assessments are explored from STEM labor workforce, scientific literacy, and Florida educational leadership perspectives.

Assessment Results

Results of the 2015 Program for International Student Assessment (PISA), the most recent PISA to emphasize science, put U.S. 15-year-old students in the middle of the pack among 71 participating countries (Kastberg et al., 2016). These students ranked 17^{th} in science among the 35 OECD countries, and 24^{th} among all participating countries (Desilver, 2017; Kastberg et al., 2016). These positions were unchanged from the 2012 PISA (Kastberg et al., 2016). Of the 5,700 U.S. students assessed on the 2015 PISA, more than 20 percent did not reach the minimum level of proficiency in science, and another 25 percent reached only the minimum level of proficiency. The PISA mean science scores for U.S. students have remained at or under the OECD average since 2006 (Organization for Economic Cooperation and Development [OECD], 2016a). In 2012, when Florida was used as a PISA benchmark state, Florida students achieved a mean scale score for science eleven points below the U.S. mean score (National Center for Education Statistics [NCES], 2017e).

The performance of U.S. 8th graders on the TIMSS science assessments has been higher than on the PISA science assessments for the three testing cycles up to and including 2015 (NCES, 2017f). Internationally, on the TIMSS 8th grade science assessment, only seven of 48 participating countries had statistically significant higher average scores than the US (Desilver, 2017) The average science score for U.S. 8th graders on the 2015 TIMSS was 30 points higher than the international average, placing them at number eight among 57 participating countries. Seventy-five percent achieved the intermediate level of proficiency or higher (NCES, 2017f). The average science score for U.S. 8th grade students increased by 17 points from 1995 to 2015; however, the average science score for Florida 8th graders dropped 22 points from 2011 to 2015 (NCES, 2017f). The 2015 TIMSS average science score for Florida 8th graders was 22 points lower than the national average, and only 65 percent of Florida 8th graders achieved the intermediate level of proficiency or higher (NCES, 2017f).

The NAEP science assessment reports for 2009, 2011, and 2015 show that only about one-third of 8th grade students achieved the minimum level of proficiency in science (NCES, 2017c). A smaller percentage of Florida 8th grade students achieved the minimum level of proficiency for those years (NCES, 2017c). While the NAEP science achievement gap between Florida 8th grade students and the U.S. average narrowed from five percent in 2009 to only one percent in 2015, only 33 percent of Florida students achieved proficiency in 2015 (NCES, 2017c).

The 8th grade FCAT 2.0 Science/SSA results show that less than half of students have ever achieved the minimum level of proficiency in science since the state began standardized assessment of science in 2007 (Florida Department of Education [FLDOE], 2012b, 2015e, 2016c). Students who achieved the minimum level of proficiency increased from 38 percent in 2007 to 49 percent in 2014, and dropped to 48 percent in 2015, 2016, and 2017 (FLDOE, 2017j). The results of the PISA, TIMSS, NAEP, and 2007–2017 8th grade FCAT 2.0 Science/SSA are summarized in Figure 2.

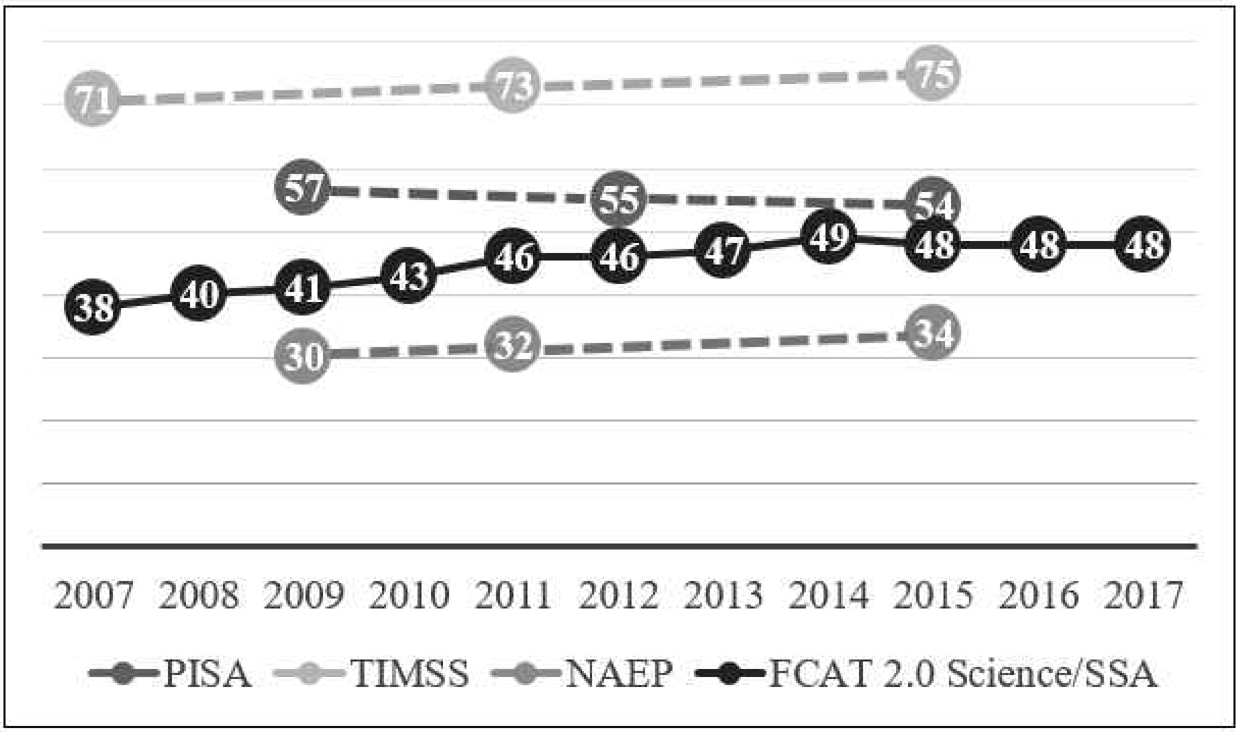

Figure 2. Percent of students, minimum proficiency or higher, 2007–2017.
Dashed lines represent years in which science assessments were not conducted. PISA, TIMSS, and NAEP data compiled from NCES Data Explorer (NCES, 2017e). FCAT 2.0 Science/SSA data for Florida 8th grade students compiled from Florida Standards Assessments Science Results website (FLDOE, 2016c).

Implications

On a national level, there is little agreement about the implications of the science assessment score trends or the ranking of the US among other participating nations. The international rankings and stagnation of science achievement of U.S. students have been deemed a national crisis affecting the U.S. STEM labor workforce pipeline by some researchers, authors, and politicians (Duncan, 2010; Langdon et al., 2011; National Academy of Sciences [NAS], 2007; National Science Board [NSB], 2007; U.S. Department of Education [U.S. ED], 2016b, 2016a). This viewpoint is challenged as a myth, manufactured crisis, or a simple lack of scientific literacy by others (Anelli, 2011; B. A. Brown, Reveles, & Kelly, 2005; Charette, 2013; Faulkner, 2012; Metcalf, 2010; Salzman, 2013; Wenning, 2007).

STEM Workforce Pipeline Perspective

In the 1990s, the American Association for the Advancement of Science (AAAS) and the National Research Council concluded that U.S. science instruction is neither suitable nor sufficient to equip students with the scientific knowledge and skills of today's technological world (American Association for the Advancement of Science [AAAS], 1993; National Research Council [NRC], 1996). In 2007, the National Science Board (NSB) declared careers and education in science, technology, engineering, and math (STEM) a national priority (NSB, 2007). In 2010, U.S. Secretary of Education Arne Duncan stated:

> Unfortunately, the 2009 PISA results show that American students are poorly prepared to compete in today's knowledge economy…Americans need to wake up to this educational reality—instead of napping at the wheel while emerging competitors prepare their students for economic leadership. (Duncan, 2010, paras. 6, 11)

From 2001 to 2011, the growth in STEM jobs grew three times as fast as growth in non-STEM jobs, and U.S. businesses often experienced difficulty in filling those jobs (Langdon et al., 2011). In 2012, the President's Council of Advisors on Science and Technology issued a report to President Obama stating that the US must produce a million more STEM professionals by

2022 if the nation is to "retain its historical preeminence in science and technology" (Olson & Riordan, 2012, p. i). In 2016, the NSF predicted that by the year 2020, occupations requiring some type of post-secondary STEM education will increase by 25 percent (National Science Foundation [NSF], 2016). That same year, the U.S. Department of Education (U.S. ED) projected that by the year 2020, all occupations requiring skills in STEM, degreed and non-degreed, will grow more than 65 percent, while the US has an inadequate number of students becoming proficient in STEM skills to meet the projected demand (U.S. ED, 2016b).

Other researchers, institutions, and authors view the STEM workforce pipeline shortage as a myth or a manufactured crisis. A 2015 report by the National Student Clearinghouse Research Center showed that, from 2009 to 2014, the number of STEM bachelor's degrees granted in the US grew by 19 percent (National Student Clearinghouse, 2015). Anft (2013) stated that there is actually significant unemployment among STEM professionals, and that the crisis has been created by technology companies and foundations seeking government funding. A 2014 report by the U.S. Census Bureau, based on 2010 census data, showed that 74 percent of those with a bachelor's degree in STEM were not employed in STEM occupations (U.S. Census Bureau, 2014). Xue and Larson (2015), and Metcalf (2010), assert that the term STEM is too broadly defined in the U.S. Department of Labor's STEM workforce projections, and that there is no agreement as to whether the term applies only to degreed STEM professionals or more broadly to include lab technicians and skilled trades. The 2012 science and engineering indicators report by the NSB states that because of the lack of agreement on the definition of STEM, "projections of employment growth are plagued by uncertain assumptions and notoriously difficult to make" (p. 3-12). Charette (2013) noted this NSB statement as support for his assertion that the U.S. STEM labor shortage is a myth:

> Even as the Great Recession slowly recedes, STEM workers at every stage of the career pipeline, from freshly minted grads to mid- and late-career PhDs, still struggle to find employment as many companies, including Boeing, IBM, and Symantec, continue to lay off thousands of STEM workers. (p. 46)

If the projected STEM workforce pipeline shortage and lack of scientific literacy do exist, neither is due to a lack of student interest in science and STEM careers (ACT, Inc., 2016; Langdon et al., 2011; OECD, 2015c; U.S. ED, 2016b). In 2015, students participating in PISA were surveyed about their attitudes toward science and their expectations for pursuing science-based careers (OECD, 2015c). This survey showed that 38 percent of the U.S. 15-year-old students surveyed expected to work in science-related careers by the age of 30. Of the U.S. high school students who took the ACT in 2015, 46 percent expressed interest in pursuing STEM majors in college, but only 20 percent met the ACT's STEM benchmark (ACT, Inc., 2016). The NCES reported that, from 2003 to 2009, 28 percent of bachelor's degree students and 20 percent of associate's degree students chose a STEM major (Chen & Soldner, 2013). The attrition rate for STEM majors from 2003 to 2009 was 48 percent for bachelor's and 69 percent for associate's degree students (Chen & Soldner, 2013; NSB, 2016).

Some researchers view the rankings of U.S. students on international assessments as misleading (Carnoy, 2015; Carnoy & Rothstein, 2013; Charette, 2013; Gibbs & Fox, 1999; Metcalf, 2010). Carnoy and Rothstein (2013), researchers for the National Education Policy Center, state that because the US has a much higher proportion of low SES students than most countries that participate in PISA and TIMSS, the average performance of U.S. students appears low when compared to all other participating countries. When compared only to three top-performing countries (Canada, Finland, and South Korea) and three “similar post-industrial countries” (p. 10) (France, Germany, and the United Kingdom), and controlled for socio-economic status, the U.S. rankings increase (Carnoy & Rothstein, 2013).

In a *Washington Post* interview, Marc Tucker, president and CEO of the National Center on Education and the Economy, criticized Carnoy and Rothstein’s (2013) study on two points (Strauss, 2013). First, the US cannot choose the countries with which it competes in a global economy, rendering the argument irrelevant. Second, because socio-economic status is a much better predictor of academic success in the US than in other countries, the assessment results show only that the US does less than other countries to educate low-SES students (Strauss, 2013). The purpose of international assessments such as PISA and TIMSS is not to provide an arena for international competition, but to provide a way for nations to improve their educational systems by learning from top-performing countries (Gibbs & Fox, 1999; Sahlberg & Hargreaves, 2011; Tucker, 2011).

Scientific Literacy Perspective

Other researchers state that the trends in science assessment results indicate not a need for an increase in the STEM career pipeline, but rather a need for a higher level of science or scientific literacy among all students (AAAS, 1989; DeBoer, 1991, 2000, 2002, Dewey, 1916, 1916; Holbrook & Rannikmae, 2009; Hurd, 1958, 1982, 1986, 2000; NGSS Lead States, 2013b; OECD, 2016d; Roth & Barton, 2004; Wenning, 2007; Wong & Pugh, 2001, 2001). However, as with the disagreement over the significance of standardized science assessment results, there is disagreement over the definition of the term scientific literacy (Holbrook & Rannikmae, 2009). Definitions have changed over time and vary among countries, organizations, and individuals. A variation of this term, science literacy, was coined by Paul DeHart Hurd in 1958 (Hurd, 1958), but the concept goes back to John Dewey's time.

Prior to the 1900s, science education was viewed as irrelevant to cultured, classically-educated, men (Huxley, 1881; Norton, 2001). "Scientific education was despised by practical business men because it seemed not only unnecessary, but actually harmful as a preparation for business" (Norton, 2001, para. 4). Science was studied only by those men deemed able to understand its complexity and use it to expand the body of scientific knowledge (Atkin & Black, 2003, 2010; J. M. Bower, 2005; Brooks & Brooks, 1999; Bybee, 1995, 1997; Champagne, 1997; Ornstein & Levine, 2000). Thomas Huxley argued that, "scientific education is every bit as culturally valuable as a humanities-based education, if not more so" (Huxley, 1881, p. 7). Edward Youmans, who published *The Handbook of Household Science* in 1859 and founded *Popular Science Monthly* in 1872, advocated a move away from the esoteric treatment of science toward its more practical applications in everyday life (1867). In the early 1900s, John Dewey continued this advocacy of science education for all as essential to the economic, social, and cultural development of the nation (DeBoer, 1991, 2000, Dewey, 1902, 1910, 1916, 1938; Wong & Pugh, 2001).

Today's definitions of scientific literacy range in focus from science content, to science practices, science appreciation and wonder, science practicality, and scientific critical thinking (Holbrook & Rannikmae, 2009). The definition of scientific literacy used in the National

Science Education Standards (upon which the NAEP science assessment is based) is, "the knowledge and understanding of scientific concepts and processes required for personal decision making, participation in civic and cultural affairs, and economic productivity" (NAS, 1996, p. 22). The OECD definition (upon which the PISA science assessment is based) is, "the ability to engage with science-related issues, and with the ideas of science, as a reflective citizen, explain phenomena scientifically, evaluate and design scientific inquiry, and interpret data and evidence scientifically" (OECD, 2016d, p. 1). Regardless its definition, the science assessment results of PISA, TIMSS, NAEP, and FCAT 2.0/SSA, all of which purport to assess science literacy, indicate that many American students are not science-literate (AAAS, 1989, 1993; Anelli, 2011; Bohrnstedt, 2016; Charette, 2013; FLDOE, 2017p; Kastberg et al., 2016).

Florida Educational Leadership Perspective

For Florida educational leaders, the FCAT 2.0 Science/SSA results are significant for at least three reasons. First, the Every Student Succeeds Act (ESSA) of 2015 requires states to assess science at least three times from grades 3 through 12 (ESSA, 2015). Second, educational leadership has been shown to have a significant impact on student achievement (Waters, Marzano, & McNulty, 2003). Third, results of FCAT 2.0 Science/SSA impact the achievement of the state's educational goals, as well as school and school district evaluations (§1008.31 Fla. Stat., 2016).

Like NCLB, ESSA requires states to assess science achievement of at least 95 percent of students, at least once in grades 3–6, 6–9 and 10–12 (ESSA, 2015). Unlike NCLB, ESSA allows states flexibility in determining how these assessments take place (ESSA, 2015). States may opt to use smaller assessments spread throughout the school year rather than a single end-of-year assessment. Instead of state-created assessments, states may opt to use nationally recognized assessments such as ACT or SAT (Barnaby, 2017). Educational leaders at all levels are responsible for making these assessment decisions and complying with federal law (Barnaby, 2017).

A 2003 meta-analysis of over 5,000 studies on the effects of leadership on student achievement found an effect size of 0.25 between leadership and student achievement (Waters, Marzano, & McNulty, 2003). Leaders can have a positive impact on achievement, and conversely, leaders can have a negative impact on student achievement when they concentrate on the wrong practices or misjudge the impact of changes they try to implement in schools (Waters et al., 2003). Their findings that student achievement was an average of 10 percentile points higher in schools with educational leaders who focused on 21 leadership responsibilities, including culture, order, discipline, and relationships, than in schools with leaders who did not focus on these responsibilities (Waters et al., 2003). These findings are supported by Hattie (2009), in his synthesis of meta-analyses on principals and school leaders. Hattie (2009) differentiated educational leadership into two categories—transformational and instructional—and found that educational leaders who focused on instructional leadership had an effect size of 0.36 on student achievement.

Florida law and the Florida State Board of Education's strategic plan state that the top goal of the state's K-20 education system is "highest student achievement, as indicated by evidence of student learning gains at all levels" (§1008.31 Fla. Stat., 2016; FLDOE, 2014f, para. 2). For 8th grade science, this goal is not being met (FLDOE, 2015f, 2016c, 2017c). Florida law also makes FCAT 2.0 Science/SSA one of the components of the state's school and school district grading model (§1008.34 Fla. Stat., 2016; Fla. Admin. Code R. 6A-1.09981, 2016; FLDOE, 2016a). The 8th grade FCAT 2.0 Science/SSA overall pass rate (percentage of students who reached Achievement Level 3 or higher) counts for up to 100 points in each middle school grade, and up to 100 points times the number of middle schools in each school district grade (FLDOE, 2017b). This pass rate is for all 8th grade students, including low SES students and ELs, regardless of the time enrolled in the school (FLDOE, 2017b). Improving the level of science achievement on the SSA is one way to help achieve the state's top educational goal and improve school and school district evaluations. One option school district leaders may choose to help improve science achievement, which is the focus of this study, is to change the type of science courses offered (FLDOE, 2015a, 2017b).

This section of the literature review has shown the status of science education as measured by PISA, TIMSS, NAEP, and FCAT 2.0/SSA. Implications of the results for the U.S. STEM workforce pipeline, scientific literacy, and Florida educational leadership were explored. The next section, science assessments, looks more closely at PISA, TIMSS, NAEP, and FCAT 2.0 Science/SSA, and explores the impacts of standardized assessments on students, teachers, and instruction. The citations used in this section are listed in Table 8.

Table 8

Science Education Literature Review Topics and Citations

Topic	Citations	
Assessment results	Desilver, D., 2017. Florida Department of Education, 2012b, 2015f, 2016c, 2017i. Kastberg, D., Chan, J. Y., & Murray, G., 2016.	National Center for Education Statistics, 2017b, 2017c, 2017e, 2017f, 2017i. Organization for Economic Cooperation and Development, 2016a.
Implications: STEM workforce perspective	AAAS, 1993. ACT, Inc., 2016. Anelli, 2011. Anft 2013. B. A. Brown, Reveles, & Kelly, 2005. Carnoy & Rothstein, 2013. Carnoy, 2015. Charette, 2013. Chen & Soldner, 2013. Duncan, 2010. Faulkner, 2012. Gibbs & Fox, 1999. Langdon et al., 2011. Olson & Riordan,2012. Metcalf, 2010. NAS, 2007.	National Science Board [NSB], 2016. National Student Clearinghouse, 2015. NRC, 1996. NSB, 2007. NSF, 2016. OECD, 2015b. Sahlberg & Hargreaves, 2011. Salzman, 2013. Strauss, 2013. Tucker, 2011. U.S. Census Bureau, 2014. U.S. ED, 2016a, 2016b Wenning, 2007. Xue and Larson, 2015.
Implications: Scientific literacy perspective	AAAS, 1989, 1993. Anelli, 2011. Atkin & Black, 2003, 2010. Bohrnstedt, 2016. Brooks & Brooks, 1999. Bybee, 1995, 1997. Champagne, 1997. Charette, 2013. DeBoer, 1991, 2000, 2002. Dewey, 1902, 1910, 1916, 1938. Youmans, 1859, 1867, 1872. FLDOE, 2017n. Holbrook & Rannikmae, 2009.	Hurd, 1958, 1982, 1986, 2000. Huxley, 1881. J. M. Bower, 2005. Kastberg et al., 2016 NAS, 1996. NGSS Lead States, 2013a. Norton, 2001. OECD, 2016c. Ornstein & Levine, 2000. Roth & Barton, 2004. Wenning, 2007. Wong & Pugh, 2001.
Implications: Florida school leadership perspective	§1008.31 Fla. Stat., 2016. §1008.34 Fla. Stat., 2016. Barnaby, 2017. ESSA, 2015. Fla. Admin. Code R. 6A-1.09981, 2016.	FLDOE, 2014g, 2015a, 2015g, 2016a, 2016c, 2017b. Hattie, 2009. Waters et al., 2003. Waters, Marzano, & McNulty, 2003.

Science Assessments

This section presents an overview and comparison of the PISA, TIMSS, NAEP, and FCAT 2.0 Science/SSA science assessments. The section concludes with an exploration of the potential impact of assessments on student performance and career expectations.

PISA

The PISA is a triennial international survey used to evaluate the educational systems of Organization for Economic Cooperation and Development (OECD) countries. The purpose of PISA is to measure the "yield of education systems", or the mastery and application of skills and competencies students have acquired near the end of their schooling (E. Scott, 2016, p. 2). The PISA measures the reading, mathematics, and scientific literacy of over 540,000 15-year-old students (usually 10^{th} grade students in the US) from 35 OECD countries and up to 36 non-OECD countries every three years, with emphasis rotating among the three subject areas each year (OECD, 2017). The PISA also collects survey data from teachers and student participants, which include the type of science course (subject-specific, integrated, or mixed) the student attended in secondary school (OECD, 2017).

Beginning in 2015, PISA was administered in a digital format, although countries could elect to use a paper-based format. The two-hour assessment contained a mixture of multiple-choice and constructed-response items (OECD, 2016b). The PISA science assessment scale scores ranged from 0 to 1,000, with cut scores for seven proficiency levels of scientific literacy: 1b, 1a, and 2 through 7, with 7 as the highest level. In addition to the student assessment, PISA includes a teacher questionnaire to collect data about science teachers' perceptions of their experience, preparation, professional development, work environment, type and subject area of science course they instruct, and the instructional methods they use (National Center for Education Statistics [NCES], 2015d).

Framework

The PISA science assessment required students to demonstrate three competencies: “1) explain phenomena scientifically; 2) evaluate and design scientific inquiry; and 3) scientifically interpret data and evidence” (OECD, 2016b, p. 20). The design of the PISA science assessment was based on content knowledge, procedural knowledge, and epistemic knowledge (OECD, 2016d). The content knowledge assessed was related to physical, life, and Earth/space sciences and science-based technology. The procedural and epistemic concepts assessed were related to the nature of science, such as the use of repeated trials and measurements to test hypotheses, control of variables, communication of results, the development of scientific knowledge through evidence-based conclusions, replication of studies, peer review, and the modification of scientific theories to incorporate new evidence (OECD, 2016b).

Standards Assessed

The 2015 PISA science assessment items were based on surveys of science standards from OECD and non-OECD countries (OECD, 2016b). Although the 2015 PISA framework did not use the term standards, the framework included statements regarding the procedural, epistemic, and content knowledge “students are likely to have acquired by the age of 15” (OECD, 2016b, p. 74). Assessment items are based on these statements. The content knowledge assessed was indicated in general topic descriptions, six each for physical, life, and Earth/space sciences. Nature of science items assessed were indicated in five general topic statements regarding procedural knowledge, and six general topic statements regarding epistemic knowledge.

The items on the PISA 2015 science assessment were distributed by type of knowledge and subject area (OECD, 2016b). The PISA 2015 science assessment emphasized physical and life sciences content over Earth/space science, and procedural knowledge over epistemic (OECD, 2016b). The PISA 2015 distribution of science assessment items, by knowledge type and subject area, is shown in Table 9.

Table 9

PISA 2015 Science Assessment: Distribution of Items

	Content area			
	Physical	Life	Earth/space	Total
Knowledge type	%	%	%	%
Content	20–24	20–24	14–18	54–66
Procedural	7–11	7–11	5–9	19–31
Epistemic	4–8	4–8	2–6	10–22
Total knowledge types	36	36	28	100

Note. Adapted from Table 2.4, *PISA 2015 Assessment and Analytical Framework: Science, Reading, Mathematic and Financial Literacy*. © 2016 by OECD Publishing. Adapted with permission.

Each item was designed to assess students' scientific literacy competencies in explaining phenomena scientifically; evaluating and designing scientific inquiry; and scientifically interpreting data and evidence (OECD, 2016b). By competency, the science items of the 2015

PISA were distributed 40–50 percent to explaining phenomena scientifically, 20–30 percent to evaluating and designing scientific inquiry, and 30–40 percent to interpreting data and evidence scientifically (OECD, 2016b). The contexts in which these items were presented included personal, local/national, and global perspectives on five general topics: 1) health and disease; 2) natural resources; 3) environmental quality; 4) hazards; and 5) the frontiers of science and technology (OECD, 2016b). The framework prescribed no distribution as to the percentage of items presented in each context.

Cognitive Demand

Beginning with the PISA 2015 science assessment, a system for determining the cognitive demand of the items was used (OECD, 2016b). Prior to 2015, only item difficulty was considered. The system used in 2015 was an adaptation of Webb's (1997) Depth of Knowledge grid, and Hess' (Hess, Jones, Carlock, & Walkup, 2009) Cognitive Rigor matrix. Instead of a two-dimensional matrix combining Bloom's Revised Taxonomy (L. Anderson & Krathwohl, 2001) and Webb's (1997) Depth of Knowledge levels, the PISA cognitive demand framework incorporated four dimensions to determine the cognitive demand level of each item. These dimensions were (a) the degree of complexity of knowledge, (b) prior knowledge, (c) cognitive operations (recall, analysis, evaluation), and (d) the level of abstraction required by each assessment item (OECD, 2016b). Based on these dimensions, items were classified as low, medium, or high cognitive demand. Low cognitive demand items comprised eight percent of the PISA 2015 science assessment and required recall or one-step procedures. Medium cognitive level items comprised 30 percent of the items and required students to use and apply conceptual knowledge. High cognitive level items comprised 61 percent of the items, and required students to analyze, synthesize, evaluate, justify, or plan in complex scenarios (OECD, 2016b).

TIMSS

The TIMSS assessment is administered every four years by the International Association for the Evaluation of Educational Achievement (IEA), an international organization of national research institutions and governmental research agencies (International Association for the Evaluation of Educational Achievement [IEA], 2017). The purpose of TIMSS, for the US, is to collect primary and middle grades educational achievement information for international comparison (E. Scott, 2016). The TIMSS assesses the science and mathematics achievement of over 580,000 4th and 8th grade students from up to 57 countries (Desilver, 2017; IEA, 2017). The most recent TIMSS took place in 2015. Unlike PISA, TIMSS is a paper-based assessment (until 2019) that assesses primarily academic content (Bohrnstedt, 2016). The paper-based (until 2019) TIMSS science assessment consists of two blocks of multiple-choice and constructed-response items, to be completed in 90 minutes (Martin, Mullis, & Foy, 2016). Student scores on the 2015 TIMSS 8th grade science assessment were divided into four achievement levels (IEA, 2016). In addition to the student assessment, TIMSS includes a teacher questionnaire to collect data about science teachers' perceptions of their experience, preparation, professional development, work environment, type and subject area of science course they instruct, and the instructional methods they use (NCES, 2015e).

Framework

Like PISA, assessment items on TIMSS are not based on the standards or curricula of any country. Representatives from participating countries decide what is important for students to know (Bohrnstedt, 2016). The TIMSS assessment items cover the three cognitive domains (knowing, applying, and reasoning) to assess students' abilities to demonstrate their knowledge in four content domains, apply what they have learned, solve problems, and reason through analysis and logical thinking (L. R. Jones, Wheeler, & Centurino, 2016).

The TIMSS science assessment is organized into overlapping content and cognitive dimensions. The content dimension specifies the subject matter to be assessed into four domains: biology, chemistry, physics, and earth science (L. R. Jones et al., 2016). The cognitive dimension specifies the thinking processes to be assessed in three domains: knowing, applying, and reasoning (L. R. Jones et al., 2016). The percentages of assessment items distributed among the content and cognitive domains are shown in Table 10.

Table 10

TIMSS 2015 8th Grade Science Assessment: Distribution of Items

Content domain	%	Cognitive domain	%
Biology	35	Knowing	35
Chemistry	20	Applying	35
Physics	25	Reasoning	30
Earth science	20		
Total	100		100

Note. Adapted from Exhibit 7, *TIMSS 2015 Science Framework*. Copyright 2016 by IEA. Adapted with permission.

Standards Assessed

The TIMSS 2015 science framework provides a list of topic areas for each content domain. Each topic area includes a list of topics, and each topic includes a list of specific objectives that represent how students should be able to demonstrate mastery of each topic (L. R. Jones et al., 2016). The numbers of topic areas, topics, and benchmarks in each content domain are shown in Table 11. In addition to these content objectives, TIMSS 2015 also assesses students' mastery of five scientific practices: (a) asking questions based on observations; (b) generating evidence; (c) working with data; (d) answering the research question; and (e) making an argument from evidence. These practices are used as the context in which some of the assessment items are framed (L. R. Jones et al., 2016).

Table 11

TIMSS 2015 8th Grade Science Topic Areas, Topics, and Objectives

	Content domain				
	Biology	Chemistry	Physics	Earth science	Totals
Number of topic areas	6	3	5	4	18
Number of topics	15	9	11	9	44
Number of objectives	36	23	34	20	113

Note. Compiled from TIMSS 2015 Science Framework.

Cognitive Demand

Unlike PISA, TIMSS 2015 does not specify a method for determining the cognitive demand or rigor in its item writing guidelines, methods and procedures manual, or assessment frameworks manual (L. R. Jones et al., 2016; Mullis, 2013; Mullis & Martin, 2013). Cognitive rigor takes into account both the cognitive skills and depth of knowledge required by assessment items (Hess et al., 2009). Instead, the TIMSS 2015 science framework manual defines three cognitive domains which are roughly equivalent to Bloom's original taxonomy of the cognitive learning domain (L. Anderson & Krathwohl, 2001; Engelhart, Furst, Hill, & Krathwohl, 1956). Items of the lowest cognitive domain, knowing, comprise 35 percent of the assessment and address students' ability to recall, recognize, and describe facts, concepts, and procedures. Items of the middle cognitive domain, applying, comprise another 35 percent of the assessment items and focus on students' ability to use knowledge to generate explanations and solve problems. Items of the highest cognitive domain, reasoning, comprise 30 percent of the assessment items and focus on students' ability to use evidence and science understanding to analyze, synthesize, and generalize in unfamiliar situations and complex contexts (L. R. Jones et al., 2016).

NAEP

The NAEP is administered by the National Center for Education Statistics (NCES) every year. The goal of NAEP is to collect educational achievement information at key stages of education across the US (E. Scott, 2016). The subject-area focus (science, reading, writing, mathematics, civics, etc.) and student sample focus (national or state focus) shift from year to year (NCES, 2017b) The three most recent NAEP science assessments were in 2009, 2011, and 2015 (NCES, 2017b). Up to 20,000 4th, 8th, and 12th grade students are assessed in national-only sample years, and up to 165,000 more students are assessed in combined national and state sample years.

The most recent NAEP science assessment was in 2015. It used both national and state samples and assessed 110,900 8th grade students (NCES, 2017b). The 2015 NAEP 8th grade science assessment was both paper-based and digital. It included selected-response, constructed-response, and combination items. Students were given 60 minutes to complete the assessment (NCES, 2017b). Subsets of the student samples were also given up to four hands-on performance or interactive computer tasks, and allowed another 30 minutes to complete these tasks (WestEd & Council of Chief State School Officers [CCSSO], 2014). The 2015 NAEP science assessment was focused on academic content in the subjects of physical, life, and Earth/space science (Bohrnstedt, 2016). Some items were framed to assess student knowledge of the nature of science and ability to identify and use the principles of scientific inquiry and technological design (WestEd & CCSSO, 2014).

Framework

Items on the 2015 NAEP science assessment were based on content area and science practice dimensions. The content dimension included physical science, life science, and Earth/space science. The science practices dimension included the areas of identifying science principles, using science principles, using scientific inquiry, and using technological design (WestEd & CCSSO, 2014). Items on the 2015 8th grade science assessment were distributed by both dimensions according to the percentage of student response time, as shown in Table 12.

Table 12

NAEP 2015 8th Grade Science Assessment: Distribution of Items

Content area	Student response time %	Science practice	Student response time %
Physical science	30	Identifying science principles	25
Life science	30	Using science principles	35
Earth/space science	40	Using scientific inquiry	30
		Using technological design	10
Total	100		100

Note. Adapted from Exhibits 17 and 18, *Science Framework for the 2015 NAEP*. In the public domain.

The most heavily weighted areas, by percentage of student response time, were the content area of Earth/space science and the science practice area of using scientific principles (WestEd & CCSSO, 2014). By response time, 50 percent of the items were selected-response, and 50 percent were constructed-response. By number of items, most of the items were traditional selected-response (WestEd & CCSSO, 2014).

The NAEP 2015 8^{th} grade science assessment scale scores ranged from zero to 500. Cut scores divide scale scores into three achievement levels: basic, proficient, and advanced. Each level contains a general description of expected performance, and detailed descriptions for expected performance in each content area and science practice at each achievement level (WestEd & CCSSO, 2014). Like PISA and TIMSS, the NAEP also includes a teacher questionnaire to collect data about science teachers' perceptions of their experience, preparation, professional development, work environment, type and subject area of science course they instruct, and the instructional methods they use (NCES, 2015e).

Standards Assessed

The *National Science Education Standards* (NSES) (NRC, 1996) and *Benchmarks for Science Literacy* (AAAS, 1993) were used to identify the science content assessed in the 2015 NAEP science assessment (WestEd & CCSSO, 2014). (The NSES were updated with the national Next Generation Science Standards [NGSS] in 2013. The NGSS were not used for the 2015 NAEP science assessment, as its development was already underway at the time of the NGSS release (NGSS Lead States, 2013b)). Content that appeared in both documents was included in the NAEP science assessment. Content that did not appear in both documents was discussed and decided upon.

These topics and subtopics were used to generate content statements and commentary that specified the knowledge to be assessed and not assessed for each subtopic. Content statements were labeled by content area (P, L, or E for physical, life, or Earth/space science), grade level, and sequential number. By content area, the 2015 NAEP 8th grade science assessment includes 16 physical science, 15 life science, and 12 Earth/space science content statements (WestEd & CCSSO, 2014).

Each of the science practices includes statements of overlapping performance expectations, which means that more than one of them could be assessed in the same item (WestEd & CCSSO, 2014). The science practices were assessed by using them as the context for the assessment of the content statements. Assessment item development was guided by descriptions of the performance expectations for each science practice, as well as descriptions of how each practice is used within each content area (WestEd & CCSSO, 2014).

Cognitive Demand

The NAEP 2015 8th grade science assessment framework specified four levels of cognitive demand (WestEd & CCSSO, 2014). These cognitive demand levels were similar to Bloom's original taxonomy of learning domains (Engelhart et al., 1956). Items of the lowest level required students to recall basic facts. Items of the second level required students to use procedural knowledge. Items of the third cognitive demand level required students to explain and predict. Items of the highest cognitive demand level required students to know when and where to apply knowledge in complex or unfamiliar scenarios (WestEd & CCSSO, 2014). The NAEP science assessment framework did not specify the distribution of items by cognitive demand (WestEd & CCSSO, 2014).

FCAT 2.0 Science/SSA

As the instrument upon which this study is based, the FCAT 2.0 Science/SSA administration, format, scoring, standards, reliability, and validity of the assessment are discussed in detail in the instrumentation section of Chapter III, Methodology. The FCAT 2.0 Science/SSA assesses the science achievement of about 400,000 Florida 4th and 8th grade students every year (FLDOE, 2017j). The purpose of FCAT 2.0 Science/SSA is to measure student achievement of Florida's Next Generation Sunshine State Standards (NGSSS) for science. The NGSSS for science were adopted in 2008, implemented in 2009, and first assessed in 2012 (FLDOE, 2017o). Students in grades 5 and 8 participate in the FCAT 2.0 Science/SSA each spring (FLDOE, 2017o). The assessment covers the subjects or content areas of nature of science, physical science, life science, and Earth/space science. The results show that less than half of 8th grade students achieved the minimum level of proficiency in science from 2007 through 2017. Unlike PISA, TIMSS, and NAEP, the FCAT 2.0 Science/SSA does not include a teacher questionnaire to collect data about science teachers' perceptions of their experience, preparation, professional development, work environment, type and subject area of science course they instruct, and the instructional methods they use (FLDOE, 2017j).

Framework

The 8th grade FCAT 2.0 Science/SSA is a paper-based test, consisting of 60 to 66 four-option, multiple choice items, administered in two 80-minute sessions (FLDOE, 2017e). These include both scored items and non-scored items used for field testing and statistical analysis (FLDOE, 2014c). The 2013 through 2017 FCAT 2.0 Science/SSA scale scores used the 2012 FCAT 2.0 Science scores as baseline (FLDOE, 2014c). Raw scores (number of items correct, by subject area) were converted to scale scores ranging from 140 to 260 points, with more points awarded for higher-complexity items. Scale scores were divided into five achievement levels, ranging from 1 (lowest) to 5 (highest). The minimum passing scale score was 203, which corresponds to Achievement Level 3 (FLDOE, 2013d, 2014h, 2015f, 2016h, 2017e).

Standards Assessed

Assessment items of FCAT 2.0 Science/SSA were based on Florida's Next Generation Sunshine State Standards (NGSSS) for science, and organized into four reporting categories of nature of science, physical science, life science, and Earth/space science (FLDOE, 2012c).

Florida's 8th grade NGSSS for science are comprised of 18 standards. The standards are broken down by content category into 109 benchmarks, 94 of which are assessed on the 8th grade FCAT 2.0/SSA (FLDOE, 2012c). The percentage of raw-score points derived from each content category of the 8th grade FCAT 2.0 Science/SSA remained constant from 2013-17 at 19 percent nature of science, and 27 percent each for physical, life, and Earth/space science (FLDOE, 2013c, 2014e, 2016f, 2017l, 2017l).

Cognitive Demand

Items on the 2013 through 2017 FCAT 2.0 Science/SSA were categorized by cognitive complexity as well as by content category (FLDOE, 2017l). Low-complexity items involved recall and recognition. Moderate-complexity items involved flexible thinking, reasoning, and problem-solving skills. High-complexity items involved analysis and abstract reasoning (FLDOE, 2017l). The 2013 through 2017 8th grade FCAT 2.0 Science/SSA derived 10 to 20 percent of points from low-complexity items, 60 to 80 percent from moderate-complexity items, and 10 to 20 percent from high-complexity items (FLDOE, 2013c, 2014e, 2016f, 2017l, 2017l).

Comparison of Science Assessments

Although each assessment uses different terminology, all assess student knowledge of science in the categories of nature of science, physical science (separated into chemistry and physics in TIMSS), life science (biology in TIMSS), and Earth/space science. While TIMSS, NAEP, and FCAT 2.0 Science/SSA are focused on academic content, PISA is designed to measure students' ability to apply reading, mathematics, and science skills and content knowledge to solve real-world problems (Bohrnstedt, 2016; FLDOE, 2012c, 2016e). The PISA also differs from the other assessments in that it assesses 15-year-old students rather than 8th grade students (generally ages 13–14) (Fleischman, Hopstock, Pelczar, & Shelley, 2010).

The PISA and TIMSS science assessments emphasize physical and life sciences while NAEP emphasizes Earth/space science (Mullis, 2013; OECD, 2016b; WestEd & CCSSO, 2014). The FCAT 2.0/SSA places equal emphasis on physical, life, and Earth/space sciences (FLDOE, 2013c, 2014e, 2016f, 2017l, 2017l). Nature of science is assessed on PISA, TIMSS, NAEP, and FCAT 2.0 Science/SSA by framing content-related items in the context of scientific inquiry (FLDOE, 2016e; L. R. Jones et al., 2016; OECD, 2016b; WestEd & CCSSO, 2014).

Only FCAT 2.0 Science /SSA is purely a multiple-choice item assessment (FLDOE, 2017e). The PISA, TIMSS, and NAEP science assessments contain both multiple-choice and constructed-response items (FLDOE, 2017l; Mullis, 2013; OECD, 2016b; WestEd & CCSSO, 2014). Only the NAEP science assessment includes hands-on performance tasks (WestEd & CCSSO, 2014). At 160 minutes, the FCAT 2.0/SSA is the longest test by time allowed for completion (FLDOE, 2017e). At 60 minutes, the NAEP science assessment is the shortest (NCES, 2017b).

Of the four assessments, only FCAT 2.0 Science/SSA is a purely standards-based assessment (FLDOE, 2017o). While the NAEP is based on the National Science Education Standards, the final determination of science content to be assessed is made by committees at the beginning of each science assessment cycle (NRC, 1996; WestEd & CCSSO, 2014). The PISA and TIMSS science assessments are not based on the standards of any country. Like NAEP, the science content to be assessed is determined by committees at the beginning of each science assessment cycle (L. R. Jones et al., 2016; OECD, 2016b).

Of the four assessments considered, the PISA science assessment places the most emphasis on items of high cognitive demand. Because the purpose, scale scores, and cognitive complexity of items differ among all four assessments, direct comparison of scores across assessments impractical (E. Scott, 2016). Ho (2007), and Carnoy and Rothstein (2013), caution against comparing trends on international , national, and state assessments. Labaree (2013) states that because, "one [PISA] measures what is relevant but not taught; the other [state assessments] measures what is taught but not relevant" (2013, p. 2), comparisons are all but impossible. Despite this, analysis of the scores and demographic data available within each assessment may reveal useful information. For example, the score trends and demographics of U.S. students on PISA, TIMSS, and NAEP show that family and community characteristics are powerful influences on student achievement, and that the achievement gap between proficient and below-average students is widening (Carnoy, 2015; Hanushek, Peterson, & Woessmann, 2014; NCES, 2007; E. Scott, 2016). Analysis of PISA score trends shows that not only does the achievement of U.S. students from low SES families lag other nations. Even U.S. students from educated, privileged families lag behind other nations (Hanushek et al., 2014).

Implications for Student Performance and Career Expectations

Some claim that high-stakes testing at the secondary level prompts teachers to cover massive amounts of information, and so moving beyond an educational experience based in anything more that rote memorization is difficult (Adler, Dougan, & Garcia, 2006; Hargrove et al., 2000; Vogler & Virtue, 2007). Many science teachers blame standardized assessment for narrowing the curriculum and preventing them from engaging students in more in-depth learning through inquiry instruction (Au, 2007; David, 2011; Hargrove et al., 2000; NRC, 1996). While disagreement exists on the usefulness of and limitations posed by international, national, and state assessments, there is evidence that high-stakes assessments may have unintended impact on students.

Although standardized assessment may help focus science instruction and increase its importance in the school curriculum, schools tend to focus on the assessment rather than the standards, causing students not to learn important but non-assessed topics (Britton & Schneider, 2010). Pellegrino et al. (2014) stated that, "most current tests do not require students to demonstrate knowledge of the integration between scientific practices and conceptual understanding" (p. 15). Students' thinking and problem-solving skills are difficult to measure with the multiple choice and short-answer items prevalent on standardized assessments (Britton & Schneider, 2010). Even when assessments are translated carefully and written to avoid cultural bias, English learners (ELs), low socio-economic status (SES) students and students from varied cultures often have difficulty with standardized assessments (Britton & Schneider, 2010; B. D. Jones, 2007; Slovacek, Whittinghill, Flenoury, & Wiseman, 2012). Visone (2009, 2010) questioned whether science assessments test science or reading skills, and found that many standardized science assessments do not take item readability into account in their validity measurements.

Standardized assessments also impact student self-efficacy and motivation. In countries that participate in PISA and other international standardized assessments, students are less likely to pursue science-based careers (Han, 2016). Although secondary school students may show an interest in STEM careers, low achievement on standardized science assessments may cause that interest to wane by the time they begin making decisions about higher education and careers.

Bandura's self-efficacy theory suggests, and ample research evidence shows, that students' low self-efficacy in science leads to low achievement on standardized assessments, diminished effort, and low propensity to pursue science-based careers (Bandura, 1997, 2000; Black & Wiliam, 2004; Britner & Pajares, 2006; Marsh, Chanal, & Sarrazin, 2006; Marsh, Trautwein, Ludtke, Koller, & Baumert, 2005; Neuenschwander, Vida, Garrett, & Eccles, 2007; Pajares, 1996). Hackett and Betz (1981) linked low self-efficacy of female students to the under-representation of women in STEM career fields. Similar links between low science self-efficacy, beginning with low achievement in the middle school years, and the under-representation of women and minorities in STEM careers were found in studies by Leslie, McClure, and Oaxaca (1998), and Wilson, Bates, Scott, Painter, and Shaffer (2015).

Even students who show an interest in science, perform well in classroom environments, and have high self-efficacy often perform poorly on standardized assessments. In annual studies study by the Texas Education Agency's Student Assessment Division, an average of 88 percent of 10th grade students pass their high school science courses, while only 63 percent pass the state's standardized assessment for those courses (Student Assessment Division, 2008). This can have a negative impact on self-efficacy. Self-efficacy can affect standardized assessment scores. Conversely, standardized assessment scores can affect students' self-efficacy, both positively and negatively (Arens et al., 2017; Aschbacher, Ing, & Tsai, 2014; Marsh et al., 2006, 2005). Self-efficacy and performance are both "determinants and consequences of each other" (Marsh et al., 2006, p. 101).

Standardized assessments may raise student achievement by encouraging students with high motivation and self-efficacy to put more effort into their work (Harlen & Deakin Crick, 2003). However, standardized assessments also may negatively impact students who perform well on those assessments. Top-performing students may be extrinsically motivated by standardized assessments. Extrinsic motivation may undermine intrinsic motivation. Without intrinsic motivation, interest in a particular subject or career path may wane when the extrinsic motivation no longer is present (DeBard & Kubow, 2002; B. D. Jones, 2007; Lepper, Greene, & Nisbett, 1973).

This section has outlined the PISA, TIMSS, NAEP, and FCAT 2.0 Science/SSA science assessments, and reviewed research on the potential impact of standardized assessment on student performance and science-related career expectations. Table 13 summarizes the topics and citations used in this section. The next section, science standards, reviews research on the influence of science standards on scientific literacy.

Table 13

Science Assessments Literature Review Topics and Citations

Topic	Citations	
PISA	Anderson, L., & Krathwohl, D. R. (Eds.), 2001. Hess, K. K., Jones, B. S., Carlock, D., & Walkup, J. R., 2009.	OECD, 2016a, 2016b, 2016c, 2017. Scott, E., 2016. Webb, N. L., 1997.
TIMSS	Anderson, L., & Krathwohl, D. R. (Eds.), 2001. Bohrnstedt, G., 2016. Desilver, D., 2017. Engelhart, M. D., Furst, E. J., Hill, W. H., & Krathwohl, D. R., 1956. Hess, K. K., Jones, B. S., Carlock, D., & Walkup, J. R., 2009. IEA, 2016, 2017.	Jones, L. R., Wheeler, G., & Centurino, V. A. S., 2016. Martin, M. O., Mullis, I. V. S., & Foy, P., 2016. Mullis, I. V. S., 2013. Mullis, I. V. S., & Martin, M. O., 2013. Scott, E., 2016.
NAEP	AAAS, 1993. Bohrnstedt, G., 2016. Engelhart, M. D., Furst, E. J., Hill, W. H., & Krathwohl, D. R., 1956. NCES, 2017b.	National Research Council, 1996. NGSS Lead States, 2013. Scott, E., 2016. WestEd, & CCSSO, 2014.
FCAT 2.0 Science/SSA	FLDOE, 2012c, 2013d, 2013e, 2014c, 2014f, 2014i, 2015g, 2016f, 2016h, 2017d, 2017i, 2017k, 2017m.	
Comparisons of science assessments	Bohrnstedt, G., 2016. Carnoy, M., 2015. Carnoy, M., & Rothstein, R., 2013. Fleischman, H. L., Hopstock, P. J., Pelczar, M. P., & Shelley, B. E., 2010. FLDOE, 2012, 2013, 2014, 2016a, 2016b, 2017a, 2017b, 2017c, Hanushek, E. A., Peterson, P. E., & Woessmann, L., 2014. Ho, 2007.	Jones, L. R., Wheeler, G., & Centurino, V. A. S., 2016. Labaree, D. F.,2013. Mullis, I. V. S., 2013. NCES, 2017. NRC, 1996. OECD, 2016. Scott, E., 2016. WestEd, & Council of Chief State School Officers, 2014.

Topic	Citations	
Implications for student performance	Arens, A. K., Pekrun, R., Murayama, K., Marsh, H. W., Lichtenfeld, S., & Hofe, R., 2017. Aschbacher, P., Ing, M., & Tsai, S., 2014. Bandura, A., 1997, 2000. Black, P., & Wiliam, D., 2004. Britner, S. L., & Pajares, F., 2006) Britton, E. D., & Schneider, S. A., 2010. DeBard, R., & Kubow, P., 2002. Hackett, G., & Betz, N. E., 1981. Han, S. W., 2016. Hargrove, T. Y., Jones, M. G., Jones, B. D., Hardin, B., Chapman, L., & Davis, M., 2000. Harlen, W., & Deakin Crick, R., 2003. Jones, B. D., 2007. Lepper, M. R., Greene, D., & Nisbett, R. E., 1973. Leslie, L. L., McClure, G. T., & Oaxaca., 1998.	Marsh, H. W., Chanal, J. P., & Sarrazin, P. G., 2006) Marsh, H. W., Trautwein, U., Ludtke, O., Koller, O., & Baumert, J., 2005. NRC, 1996. Neuenschwander, M. P., Vida, M., Garrett, J. L., & Eccles, J. S., 2007. Pajares, F., 1996. Pellegrino, J. W., Wilson, M. R., Koenig, J. A., Beatty, A. S., 2014. Slovacek, S., Whittinghill, J., Flenoury, L., & Wiseman, D., 2012. Student Assessment Division, 2008. Adler, S., Dougan, A., & Garcia, J., 2006. Visone, J. D., 2009. Visone, J. D., 2010. Vogler, K. E., & Virtue, D., 2007. Wilson, D. M., Bates, R., Scott, E. P., Painter, S. M., & Shaffer, J., 2015.

Science Standards

Even before the term scientific literacy was coined by Hurd in 1958, attempts have been made to define what students need to learn to achieve scientific literacy (AAAS, 1993; B. A. Brown et al., 2005; Holbrook & Rannikmae, 2009). The purpose of science education standards and benchmarks is to define scientific literacy by specifying what students need to know and understand to achieve scientific literacy (AAAS, 1989, 1993). Benchmarks further define the standards by specifying how students will demonstrate progress, and how that progress will be assessed (Hollweg & Hill, 2003). Science education standards also guide instruction, teacher professional development, and teacher evaluation (NRC, 1996). This section of the literature review provides a brief historical background of national and Florida science standards, and reviews research related to the consistency of science standards and their alignment with science assessments.

Historical Background

In 1990, the American Association for the Advancement of Science (AAAS) published *Science for All Americans*, followed by *Benchmarks for Science Literacy* in 1993. These publications were one of the earliest attempts to define scientific literacy for all U.S. high school graduates (AAAS, 1993, 2001, 2009; NRC, 1996; Rutherford, 1990). In 1996, the National Research Council published the first U.S. National Science Education Standards (NSES), designed as guidelines for the development of state standards (NRC, 1996; Tanner & Tanner, 1990). In 2001, the AAAS published its *Atlas of Science Literacy* (2001), which offered K–12 concept maps of each of the over 100 benchmarks in *Benchmarks for Science Literacy*. The AAAS publications and the NSES included standards for the physical, life, and Earth/space sciences, and nature of science. The publications emphasized scientific practices and integrated content across scientific disciplines (NAS, 1996).

Influenced by the No Child Left Behind Act of 2002, all states except Iowa had developed their own science education standards by 2006 (Tanner & Tanner, 1990; Wright & Sunal, 2006). Although the NSES and the AAAS publications influenced state standards, no state adopted the NSES completely, partly due to controversy about states' rights, evolution, and creationism (Branch, 2008; Brayboy, Faircloth, Lee, Maaka, & Richardson, 2015; Haught, 2008; E. C. Scott & Branch, 2008; Tanner & Tanner, 1990; Wright & Sunal, 2006). Florida first included standards for nature of science and physical, life, and Earth/space sciences in its Sunshine State Standards in 1996 (FLDOE, 2008b). In 2007, after agreeing that its science standards were “a mile wide an inch deep” (FLDOE, 2008b, p. 8), the FLDOE’s Office of Mathematics and Science formed a committee to update its science standards. The product of this committee was Florida’s Next Generation Sunshine State Standards (NGSSS) for science (FLDOE, 2008b, 2017l). The NGSSS for science were implemented in 2008, and are the basis for the assessments used for this research (FLDOE, 2017l).

In 2012, the National Research Council published *A Framework for K–12 Science Education* to initiate the process of creating new national science standards (NRC, 2012). In 2013, a consortium of 26 states, the National Science Teachers Association (NSTA), the AAAS, the NRC, and Achieve, Inc. developed and published the Next Generation Science Standards (NGSS), based on the NRC’s framework (NGSS Lead States, 2013b). The NGSS were needed, according to the NRC, due to advances in science and science education since the publishing of the NSES 15 years prior; shortages in the STEM workforce pipeline; and a lack of science literacy and college readiness for STEM degrees among U.S. students (NRC, 2013c).

The NGSS consist of core content ideas from the subject areas of physical, life, and Earth/space sciences; science and engineering practices; and science concepts that span subject-area boundaries (NGSS Lead States, 2013b). The NRC framework and NGSS, “highlight the power of integrating understanding the ideas of science with engagement in the practices of science and are designed to build students’ proficiency and appreciation for science over multiple years of school” (NGSS Lead States, 2013b). One of the major differences between NGSS and the older NSES is the inclusion of standards for technology and engineering practices (NRC, 2012; NGSS Lead States, 2013b). Florida’s NGSSS for science do not include standards for these areas (Florida State University [FSU], 2017).

Consistency and Alignment of Standards

To be effective on a national level, science standards must be standard; however, there is no consensus among the states on science education standards or the definition of science literacy (Lerner et al., 2012). As of December 2016, 18 states and the District of Cornell had adopted NGSS (NGSS Lead States, 2017). Only 11 of those states were among the 26 states involved in the development of NGSS (National Science Teachers Association [NSTA], 2016; NGSS Lead States, 2017). Florida is not among those states (NGSS Lead States, 2017). The remaining 32 states each have their own state-developed set of science standards (NGSS Lead States, 2017).

State science standards for the middle grades generally agree only on the inclusion of physical, life, and Earth/space sciences content and the nature of science, but the standards and benchmarks of each state vary (Haag & Megowan, 2015; Hoeg & Bencze, 2017; Yetter, Livengood, & Smith, 2017). In a study comparing state science standards on a single middle grades Earth/space science standard—lunar phases—Yetter, Livengood, and Smith (2017) found no agreement among all 50 states. They concluded that, if this simple standard is representative of all state science standards, there is no consistency among the states on what middle grades students should learn about science (Yetter et al., 2017).

If standards are to define scientific literacy and the path to its achievement, assessments must align to the standards (Porter, 2002; Vockley & Lang, 2009). Although ESSA requires states to assess science achievement, the assessments, including NAEP, are not nationalized (Haag & Megowan, 2015). In an analysis of NGSS, Hoeg and Bencze (2017) found that the 2013 NGSS prioritizes measurable and reproducible performances for students to demonstrate progress. However, the 2015 NAEP—the latest to focus on science—was based on, but not fully aligned with, the 1996 National Science Education Standards (NRC, 1996) and AAAS' *Benchmarks for Science Literacy* (AAAS, 1993; Vockley & Lang, 2009). The PISA and TIMSS science assessments are not based on any single set of standards (L. R. Jones et al., 2016; OECD, 2016b).

Marx and Harris (2006), and Vogler and Virtue (2007), found that the NCLB requirement for states to develop standards and assessments led states to develop fact-based, easy-to-measure standards, rather than standards based on skills such as deep thinking and conceptual understanding of concepts, which are more difficult-to-measure. This could help explain the low performance of U.S. students on assessments such as PISA, which focuses on higher-level thinking skills rather than memorization of content (Labaree, 2013).

Misalignment of standards to assessments and emphasis on lower-level items does not explain the stagnation in achievement of Florida students on the 8th grade FCAT 2.0 Science/SSA; however. There are 109 Florida middle grades science benchmarks, of which 94 were assessed on FCAT 2.0 Science/SSA (FSU, 2017). These standards are classified into three levels of content complexity. Twenty percent are Level 1 benchmarks, which require students to recall basic facts. Sixty percent are Level 2 benchmarks, which require students to demonstrate basic application of skills and concepts. Twenty percent are Level 3 benchmarks, which require students to demonstrate strategic thinking and complex reasoning. All are classified into one of four reporting categories (nature of science, physical science, life science, or Earth/space science) (FSU, 2017). Ten to twenty percent of the FCAT 2.0 Science/SSA assessment items are based on Level 3 benchmarks. Sixty to eighty percent are based on Level 2 benchmarks, and ten to twenty percent on Level 1 benchmarks (FLDOE, 2017l). The FLDOE provides content focus reports and test item specifications detailing the alignment of the assessment with standards and benchmarks, the contexts in which each standard and benchmark may be assessed, and clarifications on the specific content assessed and not assessed in each standard and benchmark (FLDOE, 2012c). The NGSSS for science, content focus reports, and test item specifications do not provide specific guidance as to the connections students are expected to make among the subject areas (FSU, 2017). The instrumentation section of Chapter 3 explains in more detail the alignment of Florida's NGSSS for science and FCAT 2.0 Science/SSA.

This section of the literature review has focused on historical background of national and Florida science standards, and research related to the consistency of science standards and their alignment with science assessments. Except for the 18 states and the District of Columbia that have adopted NGSS, there is little agreement among the states as to what students should learn to become scientifically literate. National and international science assessments are not aligned with NGSS or any state science standards. Florida's FCAT 2.0 Science/SSA, while subject/content-focused, aligns with Florida's NGSSS for science. A summary of the topics and citations used in this section is shown in Table 14. The next section focuses on research related three specific challenges of science education.

Table 14

Science Standards Literature Review Topics and Citations

Topic	Citations	
Historical background	AAAS, 1993, 2001, 2009. Branch, 2008. Brayboy, Faircloth, Lee, Maaka, & Richardson, 2015. FLDOE, 2008b, 2017k. Florida State University, 2017. Haught, B., 2008.	NAS, 1996. NRC, 1996, 2012, 2013b. NGSS Lead States, 2013a, 2013b. Rutherford, J. F.,1990. Scott & Branch, 2008. Tanner & Tanner, 1990. Wright & Sunal, 2006.
Consistency and alignment of standards	AAAS,1993. FLDOE, 2012, 2017. Haag & Megowan, 2015. Hoeg & Bencze, 2017. Jones, Wheeler, & Centurino, 2016. Labaree, 2013. Lerner, Goddenough, Lynch, Schwartz, Schwartz, & Gross, 2012. Marx & Harris, 2006.	NRC, 1996. NSTA, 2016. NGSS Lead States, 2017. OECD, 2016. Porter, 2002. Vockley & Lang, 2009. Vogler & Virtue, 2007. Yetter, Livengood, & Smith, 2017.

Science Education Challenges

This section of the literature review presents research related to the challenges of science education for large school districts, low socio-economic status (SES) students, and English learners (ELs). These challenges are the basis for Research Questions 2, 3, and 4 of this study.

Large School Districts

The debate over school district size dates back to the 1800s, when schools were highly localized (Boser, 2013). Consolidation of school districts reduced the number of U.S. public school districts from 117,108 in 1940 to 13,601 in 2015 (NCES, 2017a). As of 2015, large school districts (defined by the NCES as those with enrollment of 25,000 or more students) constituted only 2.1 percent of all U.S. school districts, but enrolled 35.7 percent of all U.S. public school students (NCES, 2017a). In 2015, 29 U.S. school districts had enrollments of over 100,000 students. Seven of these were in Florida (NCES, 2017a). As of 2017, eight Florida school districts have enrollments of over 100,000 students (FLDOE, 2017i). Of Florida's 67 school districts, these eight enroll 56 percent of Florida's 2.8 million public school students (FLDOE, 2017i).

Large school districts have many advantages over smaller school districts due to economies of scale. These advantages include lower per capita student production costs, greater capacity to retrieve and use information, greater access to external assistance, broader course offerings, and greater ability to offer specialized assistance to students with special needs (Boser, 2013; Hannaway & Kimball, 1998). However, large school district size also poses some challenges and diseconomies of scale (Boser, 2013; Bouck, 2004).

The size of an organization influences the choice of organizational structure, operational conditions, motivation and incentives of organizational members, and organizational outcomes (Amah et al., 2013). As school districts become larger, command, control, communication, and coordination problems often reduce accountability and organizational effectiveness and agility (Amah et al., 2013; Borman & Kimball, 2005; Boser, 2013; Davis, 2003; Driscoll et al., 2003; Hannaway & Kimball, 1998; Johnson, Kardos, Kauffman, Liu, & Donaldson, 2004; Koran, 2016). School-district-level decisions in large school districts limit local schools' autonomy, innovation, and ability to meet the diverse needs of students (Driscoll et al., 2003). Large school districts are less responsive to the people they serve, and less agile in adapting to changing student enrollment (Driscoll et al., 2003; Koran, 2016). Parents often find it difficult to make their concerns known in large school districts (Driscoll et al., 2003; Koran, 2016). Large school districts often reassign teachers among schools well into each school year, while small school districts find it easier to have the right number of teachers in each school. Teachers with general science licensure for all middle grades, rather than subject-specific licensure, are more easily reassigned in response to changing enrollment (Driscoll et al., 2003; Koran, 2016).

Large school district size has been shown to have a negative impact on student achievement even after controlling for the higher concentration of low SES students (Amah et al., 2013; Driscoll et al., 2003; Lippman et al., 1996). This negative impact is greatest in elementary and middle schools, where the achievement gaps between advantaged and low SES students, and low achieving and high achieving students, are larger in larger schools and school districts (Lippman et al., 1996; McMillen, 2004; Ruby, 2006; Tal, Krajcik, & Blumenfeld, 2006). Driscoll et al (2003) showed that population density, not just school district size, is the primary driver of larger achievement gaps in large school districts.

During periods of change, such as educational reform or new curriculum implementation, large school districts face challenges in providing adequate professional development for large teaching staffs and controlling the quality and implementation of instruction (Bybee, 1995; Drits-Esser & Stark, 2015; Gao & Wang, 2014; Pringle, Mesa, & Hayes, 2017; Sturges, 1976; Trevino Jr, Braley, Brown, & Slate, 2008). Funding for curricular materials, teacher licensure, and professional development is often lacking in large school districts with stretched budgets than in smaller, more affluent school districts. (Borman & Kimball, 2005; Boser, 2013; Bouck, 2004; Fulton, 2012; Johnson et al., 2004; McLaughlin, 2014; Rivera Maulucci, 2010; Rivera Maulucci, Brown, Grey, & Sullivan, 2014). Without professional development and support, teachers will not have an expectation of success, and without this expectation, any curriculum change is unlikely to succeed (Abrami, Poulsen, & Chambers, 2004). Teacher preparation is necessary for traditional, pre-service teachers as well as the growing number of alternatively-certified teachers entering the teaching profession from other careers or after earning a non-education degree (Sass, 2013; Woods, 2016).

Along with professional development, teacher attitudes strongly influence the quality and fidelity of implementation of new courses or curricula. Some teachers are receptive to and enthusiastic about educational innovations and curriculum change, while others resist it and continue using teaching practices with which they are comfortable (Abrami et al., 2004; Lam, Wing-yi Cheng, & Choy, 2010; Roehrig, Kruse, & Kern, 2007). Large school districts with diverse levels of experience, education, and attitudes among teaching staffs are more susceptible to these influences than small school districts with more homogeneous teaching staffs (Roehrig et al., 2007). In these school districts, science teachers often perceive that the subject of science is marginalized in the school curriculum. These challenges affect the course offerings, experiences, and outcomes of students (Bouck, 2004), and create a "support gap" for teachers in large school districts (Johnson et al., 2004, p. 2). This support gap leads to newer, less-experienced, less-qualified teachers being assigned to classrooms with the highest concentrations of low achieving students (Borman & Kimball, 2005).

Low Socio-economic Status Students

Despite continuing efforts at improving academic achievement of students of low socio-economic status (SES) students, the standardized assessment achievement gap between low SES students and others is tenacious (Becker & Luthar, 2002; Hanushek, 2010; Hattie, 2009; Ladd, 2012; Miller et al., 2013; Wiseman, 2012). The 2009, 2011, and 2015 NAEP 8th grade science assessment scores for economically disadvantaged students were an average of 27 points lower than scores for non-economically disadvantaged students (NCES, 2015a). On the 8th grade FCAT 2.0 Science/SSA for 2013 to 2017, the average percent of students who achieved the minimum level of proficiency was 37 percent for low SES students and 65 percent for higher SES students (FLDOE, 2017i). This achievement gap is not unique to large urban school districts (Bouck, 2004; Reardon, 2013). Nor is it unique to the US. Wiseman (2012) studied the effects of poverty on science achievement across 40 countries and concluded that student poverty is a strong predictor of low science achievement at every level of national economic development. Socioeconomic factors, more so than race, ethnicity, or immigrant status, were more significant in explaining differences in educational achievement, according to a 2004 RAND Corporation study (Lara-Cinisomo et al., 2004).

Many small-scale studies address the effects of specific intervention programs on the science achievement of low SES students. Classrooms with higher concentrations of minority, low SES, and low-achieving students were found more likely to be taught by less-qualified teachers with lower evaluation scores (Thadani, Cook, Griffis, Wise, & Blakey, 2010). In a five-year study of an intensive professional development program for middle school science teachers in high-poverty schools, the science achievement was significantly higher of low SES students of teachers who participated in the program than for comparable students whose teachers did not participate (Matkins et al., 2014). Low SES middle school students in Philadelphia had significantly higher science achievement than those in three matched control schools and 23 comparable middle schools when their teachers participated in a program that addressed the lack of materials, professional development, and support (Ruby, 2006). Similar results were found for low SES middle school students in Texas whose science teachers received professional development on lesson planning, inquiry instruction, and science vocabulary instruction (Jackson & Ash, 2012; Thadani et al., 2010). According to a case study by the NGSS Lead states, purveyors of the national Next Generation Science Standards, project-based, comprehensive science courses help all students, particularly those of low socioeconomic status, "recognize science as relevant to their lives and future, deepen their understanding of science concepts, and develop agency in science" (NGSS Lead States, 2013a).

Participation in STEM club activities resulted in significantly improved science achievement of low SES students (Gottfried & Williams, 2013). Learning benefits were found for low SES students in two schools that provided professional development and curriculum-based interventions (Thadani et al., 2010). A case study of a 6^{th} grade science teacher showed greater student science achievement when the teacher was trained to "engage in teaching science for social justice by taking an anti-deficit stance toward his students, expand the roles students play in science class by providing ample opportunities for them to negotiate their participation, and share authority with students by giving them the freedom to assemble a personal science portfolio" (Tan & Barton, 2010, p. 38). In Detroit, two cohorts of low SES and minority 7th and 8th graders who participated in a science intervention program using inquiry-based science units had significantly higher pass rates on the statewide test. The program was supported by aligned professional development, instructional materials, and learning technologies (Geier et al., 2008).

While there is no consensus on the specific instructional methods employed in these intervention studies, Gao (2014) found that low-performing students responded differently to different science teaching practices. For low-performing US students, traditional didactic instructional methods were significantly negatively correlated to lower science achievement, however inquiry-based instructional methods did not impact the science achievement of these students at all (Gao, 2014). Gao (2014) suggested that students of varying cultural backgrounds require different instructional approaches to meet their needs.

Regardless of the specific instructional methods, research supports the need for teacher professional development and continuing support in addressing the science achievement low SES students. Broader studies beyond the subject of science education also show that teacher preparation and support is key for the success of low SES students (Kahle, Meece, & Scantlebury, 2000; Lippman et al., 1996; McLaughlin, 2014; Ruby, 2006; Schaffer, White, & Brown, 2016; Tal et al., 2006; White et al., 2017). (Gao, 2014; Gao & Wang, 2016)

English Learners

As with the achievement gap for low SES students, the achievement gap for English learners (ELs) is persistent and even wider. On the 2015 NAEP, the average score for 8^{th} grade ELs in science was 44 points lower than for non-ELs (NCES, 2017c). From 2012 through 2017, Florida SSA/FCAT 2.0 Science scores for EL students were an average of 22 points lower than scores for non-EL students (FLDOE, 2017i). For the same time period, only 10 percent of ELs achieved the minimum passing score, as contrasted to 50.5 percent of non-ELs (FLDOE, 2017i). Confounding the problem is the fact that ELs are often minorities from low SES families (Cosentino de Cohen et al., 2005). Again, many small-scale studies address specific intervention programs on the science achievement of ELs, and that science teachers of ELs require professional development to help ELs develop language skills while learning scientific concepts and processes (Amaral et al., 2002; Fathman, Kessler, & Quinn, 1992; Santau, Secada, Maerten-Rivera, Cone, & Lee, 2010).

Amaral, Garrison, and Klentschy (2002) found ELs who participated in a four year science intervention program, which included extensive professional development training for participating teachers, had significantly higher standardized assessment scores every year they remained in the program, not only in science, but in writing, mathematics and reading as well. Maerten-Rivera, Ahn, Lanier, Diaz, and Lee (2016) implemented a curricular and professional development intervention for EL science achievement in 31 treatment schools, with 32 schools used as a control group, over a three-year period. In years two and three, the state science assessment scores for ELs in the treatment schools were significantly higher than of ELs in the control schools (Maerten-Rivera et al., 2016).

A quasi-experimental intervention designed to improve science and English reading achievement of middle school ELs resulted in significantly higher standardized assessment scores for 166 treatment students than 80 comparison students (Lara-Alecio et al., 2012). This intervention included on-going teacher professional development in inquiry science instruction methods. Treatment students also received special instruction in science vocabulary, reading, and writing. Unique components of this intervention not mentioned in the other studies reviewed in this section were the inclusion of take-home science activities and university scientist mentoring for each treatment student (Lara-Alecio et al., 2012)

Santau, Secada, Maerten-Rivera, Cone, and Lee (2010) examined elementary teachers' knowledge and practices in teaching science while supporting the needs of ELs before and after a reform-oriented professional development intervention in urban schools. Prior to the intervention, teachers felt unprepared and unsupported in teaching science to ELs, and student achievement was low. After the intervention, teachers felt they had learned how to help ELs learn science while developing their English language skills, and the science assessment scores for ELs improved significantly (Santau et al., 2010).

Using existing data from several locations across the US, Abedi (2002) found evidence showing that impact of impact of low language proficiency on assessment scores for ELs in content areas with higher language demand. Others have questioned whether standardized assessment scores for ELs are reliable or valid, even when ELs are given accommodations (Abedi, 2003; Kieffer, Lesaux, Rivera, & Francis, 2009). In a meta-analysis of 11 studies on EL accommodations in standardized assessments, Kieffer, Lesaux, Rivera, and Francis (2009) found that only one accommodation, a translation dictionary, had a measurable impact on the EL performance on assessments, and only a 0.146 effect size on the reduction of the EL/non EL achievement gap.

Some studies have shown that ELs respond well to inquiry-based science instructional methods that rely less on English language proficiency (Amaral et al., 2002; Lee, 2005; Maerten-Rivera et al., 2016; Stoddart, Pinal, Latzke, & Canaday, 2002). However, other studies have found that the relationship between inquiry-based science instruction and student achievement across demographic groups and achievement levels is not firmly established (Gao, 2014; Gao & Wang, 2016). In a study of TIMSS 2011 for 8^{th} grade students in Singapore, Chinese Taipei, and the US, Gao (2014) found that the inquiry-based instructional practices measured in her study had no significant effect on the achievement of low performing students in Singapore and the US In a study of TIMSS 2007 data for U.S. 8^{th} grade students, Gao and Wang (2016) found that “more inquiry-based instruction was not significantly associated with content and problem solving achievements across Caucasian, African-American, and Hispanic American students” (2016, p. 5404). These studies suggest that students from different countries and of different achievement levels respond differently to the various science instructional methods (Gao, 2014; Gao & Wang, 2016).

The studies reviewed in this section have shown some small-scale successes using various methods to overcome the challenges of science education in large school districts, for low SES students, and for ELs. The common theme among them has been professional development for and support of the teachers who face these challenges every day (Lee & Buxton, 2013). This professional development and support must be “focused on pedagogy that guides teachers’ efforts at improvement, and is persistent, resolved, consistent, and coherent over the long haul” (DuFour, 2011, p. 162). The citations used in this section are shown in Table 15.

Table 15

Science Education Challenges Literature Review Topics and Citations

Topic	Citations	
Large school districts	Boser, 2013.	Hannaway & Kimball, 1998.
	NCES, 2017a.	Johnson, Kardos, Kauffman, Liu, & Donaldson, 2004.
	FLDOE, 2017h.	Koran, 2016.
	Hannaway & Kimball, 1998.	Lam, Wing-yi Cheng, & Choy, 2010.
	Bouck, 2004.	Lippman et al., 1996.
	Amah et al., 2013.	McLaughlin, 2014.
	Borman & Kimball, 2005.	McMillen, 2004.
	Boser, 2013.	Pringle, Mesa, & Hayes, 2017.
	Abrami, Poulsen, & Chambers, 2004.	Rivera Maulucci, 2010.
	Borman & Kimball, 2005.	Rivera Maulucci, Brown, Grey, & Sullivan, 2014.
	Boser, 2013.	Roehrig, Kruse, & Kern, 2007.
	Bouck, 2004.	Ruby, 2006.
	Bybee, 1995.	Sturges, 1976.
	Davis, 2003.	Tal, Krajcik, & Blumenfeld, 2006.
	Driscoll et al., 2003.	Trevino Jr, Braley, Brown, & Slate, 2008.
	Drits-Esser & Stark, 2015.	
	Fulton, 2012.	
	Gao & Wang, 2014.	
Low SES students	Becker & Luthar, 2002.	Ruby, 2006.
	Hanushek, 2010.	Clifford & Ash, 2012.
	Hattie, 2009.	Thadani et al., 2010.
	Ladd, 2012.	Gottfried & Williams, 2013.
	Miller et al., 2013.	Thadani et al., 2010.
	Wiseman, 2012.	Tan & Barton, 2010.
	NCES, 2015a.	Geier et al., 2008.
	FLDOE, 2017h.	Kahle, Meece, & Scantlebury, 2000.
	Bouck, 2004.	Lippman et al., 1996.
	Reardon, 2013.	McLaughlin, 2014.
	Wiseman, 2012.	Ruby, 2006.
	Thadani, Cook, Griffis, Wise, & Blakey, 2010.	Schaffer, White, & Brown, 2016.
		Tal et al., 2006.
	Matkins et al., 2015.	White et al., 2017.
ELs	Abedi, 2002, 2003.	Kieffer, Lesaux, Rivera, & Francis, 2009.
	Amaral, Garrison, and Klentschy, 2002.	Lara-Alecio et al., 2012.
		Lee & Buxton, 2013.
	Cosentino de Cohen et al., 2005.	Lee, 2005.
	DuFour, 2011.	Maerten-Rivera, Ahn, Lanier, Diaz, and Lee, 2016.
	Fathman, Kessler, & Quinn, 1992.	NCES, 2017c.
	FLDOE, 2017h.	Santau et al., 2010.
	Gao, 2014.	Santau, Secada, Maerten-Rivera, Cone, & Lee, 2010.
	Gao & Wang, 2016.	Stoddart, Pinal, Latzke, & Canaday, 2002.

Science Courses

To improve student achievement of science standards, many school districts across the US have changed from subject-specific to comprehensive or integrated science courses in the middle grades (Banilower et al., 2013). This section begins with an explanation of the search process used to find research on comprehensive and subject-specific science instruction, of which little exists. The next section reviews research on comprehensive/integrated curricula in general. The concluding section reviews the two published articles, a conference paper, three dissertations, and a master's thesis that directly compared subject-specific science courses to integrated science courses, based on results of standardized assessments.

Literature Search Process

Systematic searches for original scientific research comparing the efficacy of comprehensive science courses to subject-specific science courses was conducted using the search terms and databases shown in Table 16. Searches focused on scholarly, peer-reviewed studies published in journals such as *International Journal of Science Education*, *Journal of Science Teacher Education, Journal of Research in Science Education, Research in Science Education, and Science Education*. These searches yielded two published, peer-reviewed, academic studies aligned with the research questions of this study—one from Switzerland, the other from the Netherlands (Åström & Karlsson, 2012; Tamassia & Frans, 2014).

Table 16

Search Terms and Databases

Databases	Search terms
EBSCOhost (Academic Search Premier, ERIC Education Source, MAS Ultra-School, ProQuest, ProQuest Middle Search Plus, ProQuest Primary Search, ProQuest Education Journals); Google Scholar; JSTOR; Omnifile; Professional Development Collection; PsycINFO; ScienceDirect; Scopus; Teacher Reference Center; UCF Libraries Catalog; Web of Science (Thomson Reuters); Wiley Online Library	Assessment; Cognitive demand; Comprehensive science; Curriculum development; Disciplinary science; Discipline-focused science; Diversity (student); Economically disadvantaged students; Educational leadership; Educational administration; Educational assessment; English learners; FCAT 2.0 science; Field-focused science; Field-specific science; Florida Comprehensive Assessment Test for science; Florida science assessment; Florida statewide science; General science; Grades 6-8/middle grades/middle school/junior high school science; Integrated curriculum; Integrated science; Interdisciplinary approach in education; Interdisciplinary science; Large school districts; Middle/junior high, multidisciplinary science; NAEP; Parental expectations; Performance based assessment; PISA; Professional development; Reciprocal effects model; School district consolidation; Science curriculum; Science education; Science education; Science education standards; Science instruction; Science learning progression; Science self-efficacy; Scientific literacy; Scientific/science literacy; Spiral science curriculum; Standardized science assessment; Standardized tests; Statewide Science Assessment; STEM attrition; STEM pipeline; STEM shortage; STEM workforce; Student evaluation; Subject-specific science; Teacher attitudes; Teaching methods; Teaching practices; Testing problems; Thematic science; Theme-based science; TIMSS; Traditional science;

Widening the search to sources other than academic journals yielded one conference paper (Faulkner, 2012), three doctoral dissertations (Alwardt, 2011; Åström, 2008; Clifford, 2016), and a master's thesis (Åström, 2007) that compared comprehensive to subject-specific science courses. The reference lists of these studies included government and organizational documents describing PISA, TIMSS, and other assessments used in the studies. Editorials regarding science interdisciplinarity (Guo, 2010; Lederman & Niess, 1997) and the integration of science with mathematics and technology were cited (Friend, 1985; Goldberg & Wagreich, 1990), but no references to empirical research comparing comprehensive to subject-specific science courses were found.

The references related to science education and integrated curriculum in the book, *Visible Learning* (Hattie, 2009), and website, “Visible Learning Plus” (Hattie, 2017) were also searched. Hattie (2009) cited only two meta-analyses related to integrated science curriculum. The first was an unpublished doctoral dissertation by Hartzler (2000), *A meta-analysis of 30 studies conducted on integrated curriculum programs and their effects on student achievement.* Of the 30 studies cited by Hartzler (2000), ten were related to science education. Of those, eight were doctoral dissertations. The two published, peer-reviewed studies related to integration of mathematics to subject-specific science.

The second meta-analysis Hattie (2009) cited was a report based on a doctoral dissertation by Hurley (2001). Hurley (2001) looked at 31 studies related to the integration of mathematics and subject-specific science, e.g. mathematics and biology, mathematics and physics, etc. Hurley (2001) included references dated from 1947 to 1998, and yielded no literature relevant to this study. Hattie (2009) included a section entitled “science programs” (p. 147), which contained references to meta-analyses of studies of science teaching strategies, which are beyond the scope of this study.

A regular search for the phrase, “comprehensive vs. subject-specific science” yielded a website called P-Reviews, administered by the School of Education at the Catholic University of Leuven, Belgium. Its “Subject Science” page invites researchers to contribute research specifically about the topic of subject-specific versus comprehensive/integrated science education; however, its reference list yielded only the previously-found international studies by

Åström (2007, 2008; 2012), Tamassia and Frans (2014), and small-scale study of 30 7th grade students in the country of Georgia (Makashvili & Slowinsky, 2009).

The scant research on comprehensive versus subject-specific science courses was noted in each of the related studies found. In 2000, the American Association for the Advancement of Science (AAAS), a proponent of comprehensive/integrated science education, stated that empirical evidence is scant supporting either a subject-specific or a comprehensive science curriculum (AAAS, 2000). In a literature review on interdisciplinary science teaching in the *Handbook of Research on Science Education*, Czerniak (2007) stated, "most of the literature on curriculum integration could be characterized as testimonials, how-to's [sic] or unit/activity ideas" (p. 544). Little changed from 2000 to 2015, when Merritt stated:

> Much of the research on the effectiveness of interdisciplinary, integrated instruction is in the form of specific case studies summarizing experiences of particular schools. There remains a paucity of research on student learning in integrated settings, and there is sparse evidence to show that students have the ability to put ideas from different courses together. (2015, para. 1)

Research related to comprehensive and integrated curricula in general, and the two published studies, four doctoral dissertations, and one masters' thesis relevant to the research questions of this study are reviewed in this section of the literature review.

Comprehensive versus Subject-specific Science Courses

A subject-specific science course presents standards and benchmarks for physical science, life science, and Earth/space science, as separate classes (Merritt, 2015). Florida school districts that offer subject-specific middle school science courses typically offer Earth/space science in 6th grade, life science in 7th grade, and physical science in 8th grade, with overarching nature of science standards and benchmarks included in each grade 6 through 8 (FLDOE, 2015a). Other terms commonly used are discipline-based, layered, field-specific, and traditional science courses (International Bureau of Education [IBE]-UNESCO, 2017a). The distinctions among these terms are not universally accepted (AAAS, 1989; IBE-UNESCO, 2017b; Stengel, 1997). The term subject-specific science course is used in this study as the subject is used in Florida's course code directory for middle/junior high (M/J) grades science courses, e.g. M/J Physical Science, M/J Life Science, and M/J Earth/Space Science (FLDOE, 2017d).

Comprehensive science courses blend standards and benchmarks for Earth/space science, life science, physical science, and nature of science. Other terms commonly used are integrated, spiral, interdisciplinary, multidisciplinary, thematic, and general science courses (Herr, 2007; NRC, 2012). There are no universally-accepted definitions among the education and scientific communities for these terms (IBE-UNESCO, 2017b; Ragel, 2015; Stengel, 1997). The term comprehensive science course is used in this study as it is used in Florida's course code directory for middle/junior high (M/J) school science courses, e.g. M/J Comprehensive Science 1, M/J Comprehensive Science 2, and M/J Comprehensive Science 3 (FLDOE, 2017a).

For this study, this terms subject-specific science course and comprehensive science course refer only to the sequencing and focus of Florida's NGSSS for science in courses offered in grades six through eight. These terms do not refer to other curriculum components such as instructional methods (inquiry, cooperative learning, direct instruction, etc.), learning experience characteristics (activity/laboratory, textbook, digital, etc.), or specific lesson design (formative/summative assessment, learning goals, differentiated instruction, etc.). The studies reviewed in this section of the literature review use different terminology, which will be explained in context.

When organized by grade level, the middle grades NGSSS for science benchmarks include benchmarks for each of the four subject areas. This provides a suggested sequencing of the standards and benchmarks for school districts that offer comprehensive science courses (FSU, 2017). While school districts are required to follow the standards and benchmarks, they are not required to follow any specific sequencing (§ 1003.41, 2016; §1003.02(1)(d), Fla. Stat., 2016). The sequencing generally follows the recommendations of the National Research Council, in *A Framework for K-12 Science Education* (2012), and the national Next Generation Science Standards (NGSS) (NGSS Lead States, 2013b). Unlike the NGSS, Florida's middle grades NGSSS for science do not provide specific guidance as to the cross-cutting concepts or connections students are expected to make among the subject areas (FSU, 2017). Another difference between NGSS and Florida's NGSSS for science is that the former includes standards for engineering practices, while Florida's NGSSS for science do not (FLDOE, 2017n).

Florida school districts that offer subject-specific middle grades science courses sequence the NGSSS for science standards by subject matter units. Physical science includes units such as force and motion; energy transformations, and structure of matter (FLDOE, 2008b). Life science includes units such as body systems, photosynthesis, and human reproduction (FLDOE, 2008b). Earth/space science includes units such as Earth's layers and landforms; rock cycle and erosion; and plate tectonics (FLDOE, 2008b). Benchmarks for nature of science are included in each subject-specific class (FLDOE, 2008b).

According to the NRC (2012), comprehensive science courses enable students to "actively engage in science and engineering practices and apply crosscutting concepts to deepen their understanding of each subject's disciplinary core ideas over multiple years of school" (NRC, 2012, p. 2). Subject-specific or discipline-based science courses, on the other hand, often neglect important historical, thematic, and non-discipline-based knowledge, according to the American Association for the Advancement of Science [AAAS] (2000). In contrast to the traditional subject-specific approach, a comprehensive curriculum is said to be a learning approach that conceives knowledge more naturally, as an indivisible whole rather than separate pieces (Silver, Strong, & Perini, 2000).

Tamassia and Frans (2014) summarized the theoretical and ideological arguments for and against comprehensive/integrated science. The most common arguments for comprehensive/ integrated science courses are that reality is not organized into separate subjects; the content is organized around real-world social problems; students are more motivated to learn the big picture rather than fragments of it; and comprehensive science supports constructivist theory (Tamassia & Frans, 2014). The most common arguments against comprehensive/integrated science are that almost no research supports it; there is no consistent definition for it; and the education and training of teachers for it present immense challenges (Tamassia & Frans, 2014). Venville, Wallace, Renne, and Malone (2002) add that comprehensive science erodes the value of science as a school subject.

Typical of the small-scale studies prevalent in the body of research on comprehensive or integrated science is an Indonesian study by Pursitasari, Nuryanti, and Rede (2015). Significant gains in student critical thinking skills were achieved by students participating in an integrated, thematic science curriculum versus a control group of students (Pursitasari et al., 2015). The title of this study reveals its major weakness: *Promoting of Thematic-Based Integrated Science Learning on the Junior High School.* Teachers in the treatment group were given training, instructional materials, and other support, while those in the control group did not. While the program may have merits, the researchers set out to promote it rather than to compare it an alternate instructional method with a group of teachers that received similar training, materials, and support.

Bing Wei (2009) conducted interview research to compare two different forms of integrated science in China. He recommended that rather than integrating the science subjects or disciplines into a general science course, each science subject instead should be integrated with subjects beyond science. The results of the study by Bing Wei showed that students who learned science with this type of integration were better able to understand the relevance of science to society and to use it to solve social problems (2009).

Other researchers have found that professional development, content knowledge, and instructional methods used to teach science are more important for student learning than the sequence in which the standards are presented (Bybee, 1997; Carr & Harris, 2001; DeBoer, 2000; Hattie, 2009; NAS, 2000; Stoica, 2015; Tyler, 1950; Zhbanova, Rule, Montgomery, & Nielsen, 2010). Teachers of comprehensive science courses must have both breadth and depth of content knowledge across the science disciplines in order to avoid inadvertently passing on or reinforcing common scientific misconceptions, and to make connections of concepts among the various subject areas (AAAS, 2011; Anelli, 2011; 2017). Santau (2010) conducted a five year study of science content knowledge of elementary school teachers across the US, and found that, while adequate, it was highly inconsistent and short of meeting the requirements for successful implementation of a comprehensive/integrated science curriculum. Teachers' knowledge of science content is directly related to the successful implementation of a comprehensive science curriculum (Herrington & Daubenmire, 2016; Santau et al., 2010).

Unlike subject-specific science curricula, the definitions, structures and methods of implementation of comprehensive science curricula vary widely, both among different states and among school districts within each state (Kesidou & Roseman, 2002). This is evidenced by the myriad terms used to describe it, such as integrated, coordinated, spiraled, interdisciplinary, multidisciplinary, thematic, and general science (Herr, 2007; Lederman & Niess, 1997; Sherriff, 2014, 2015). National publishers face challenges in writing comprehensive science curricula aligned with the myriad state science standards. In a study of nine widely-used science curricula, Kesidou and Roseman (2002) found that most used previously-written subject-specific material, re-arranged in "programs that rarely provided students with a sense of purpose for the units of study, or modeled the use of scientific knowledge so that students could apply what they learned in everyday situations" (Kesidou & Roseman, 2002, p. 522).

Hattie (2009), however, in his synthesis of meta-analyses of integrated curriculum, found that curriculum had an effect size of only 0.39 on student achievement. His view of the reason for this is not related to content knowledge. Hattie (2017) stated:

> Curriculum is among the least beneficial influences student achievement. The most powerful effects on the student relate to features within the school such as the climate of the classroom, peer influences, and the lack of disruptive students in the classroom, etc. (School Variables, FAQ 1).

The advantages and disadvantages of each type of science course may be viewed in the context of learning science principles. In Florida, as in most states, 8th grade students receive standardized assessment only once on science subject matter learned from 6th through 8th grade (FLDOE, 2017j). These assessments test student knowledge with using items of varying cognitive complexity (FLDOE, 2016e).

For assessment items of lower complexity that require recall of facts, a disadvantage of subject-specific science courses would be that students are two to three years removed from subject matter learned in the 6th grade. The learning science principle of recency states that students are more apt to remember that which they learned most recently (G. H. Bower & Hilgard, 1981; Kahana, Howard, & Polyn, 2008; Roediger, 2006). However, subject-specific courses may provide students with a more organized, coherent progression of material than comprehensive, which complies with the learning science principles of organization, coherence, and segmentation effects (Bjork, 1994; Bransford, Brown, & Cocking, 2000; Kalyuga, Chandler, & Sweller, 1999; Kozma, 2000; R. E. Mayer, 2008).

An advantage of comprehensive science courses may be that in the presentation of myriad scenarios in which a scientific concept may be applied to various scientific disciplines. This aligns with the multiple examples (Halpern & Halpern, 2005) and contiguity (R. E. Mayer, 2008) effects of learning science. However, comprehensive courses may violate the learning science principle that progression in learning is usually from the concrete to the abstract (Rutherford, 1990). “Teachers often overestimate students’ abilities to make connections of abstract ideas, and take the students’ use of the right words as evidence of learning” (Rutherford, 1990, para. 4). Comprehensive courses that focus on over-arching scientific principles may neglect the fact that, "to develop mastery, students must acquire component skills, practice integrating them, and know when to apply what they have learned" (Lovett, 2017, p. 4). Comprehensive science courses may present an advantage with the testing effect, which improves learning by repeatedly testing students on the material (Roediger, 2006). Another advantage of comprehensive courses may lie in the learning science principle of exercise (G. H. Bower & Hilgard, 1981; Kahana et al., 2008). The principle of exercise suggests students learn more, and more deeply, that which is repeated. Because students are assessed (TIMSS, NAEP, and FCAT 2.0 Science/SSA) only once at the end of 8^{th} grade on science subjects learned in the 6^{th}, 7^{th}, and 8^{th} grades, according to the principle of recency, students are likely to remember more of 8^{th} grade science than 7^{th} or 6^{th} grade science. Comprehensive science courses, which revisit important topics in different contexts throughout the middle grades, may provide students the repeated exposure to topics required for deeper learning (G. H. Bower & Hilgard, 1981; Kahana et al., 2008).

Relevant Empirical Studies

This concluding section of the literature review focuses on the seven studies found that compared comprehensive to subject-specific science courses. Only two are published, peer-reviewed, academic articles. The most recent of these is a 2014 article by Tamassia and Frans, published in the Journal of the European Teacher Education Network. The other is a 2012 study by Åström & Karlsson, published in Nordic Studies in Science Education. Another related study is from a Northeastern Educational Research Association conference paper by Faulkner, based on his 2012 doctoral dissertation. Three are doctoral dissertations: one by Clifford (2016), another by Alwardt (2011), and the third by Åström (2008). The earliest related study is a 2007 master's degree thesis, also by Åström.

The article by Tamassia and Frans (2014) provided only a literature review and commentary on the topic of integrated versus discipline-focused science courses in Flanders, Belgium. Sometime before 2014 (the paper does not state a specific year), the government of Flanders mandated a switch from subject-specific science courses to integrated science courses in non-technical secondary schools. The authors sought to find scientific evidence of the effects of integrated science on the scientific literacy of students. As with the search for literature related to this study, they found only the 2007 and 2008 studies by Åström (Tamassia & Frans, 2014). Thus, other than confirmation of the dearth of research related to comprehensive and subject-specific science, this article yielded little information useful for this study.

The 2012 study by Åström and Karlsson used three different hierarchical linear models (HLMs) to test differences in PISA 2003 science results between groups of Swedish students who had taken subject-specific science courses and those who had taken integrated science courses in secondary school. The HLMs included student variables such as gender, language spoken at home, pre-school attendance, socio-economic status, and racial/ethnic/cultural category as well as the type of science course attended. With all three models, Åström and Karlsson found that the type of science course did not show significant differences in the 2003 PISA science scores for Swedish students.

The 2008 dissertation by Åström consisted of four studies comparing integrated to subject-specific science education in Sweden. These included an expert survey, case studies of four schools, and two HLM analyses of students' science results from the 2003 and 2006 PISA

(Åström, 2008). The expert survey of 24 experts on science education and curriculum showed that the experts agreed that students can learn science in either subject-specific or integrated science courses if they are required to solve practical problems using their knowledge of science (Åström, 2008). The case studies of subject-specific and integrated science courses at two different schools showed that the primary difference was with teachers' methods of planning (Åström, 2008). Teachers of integrated science courses worked with mind maps centered around projects or themes and involved students in lesson planning. These planning methods were not found in the schools that offered subject-specific science courses (Åström, 2008). The quantitative analysis of PISA 2003 showed no difference between groups of students or between groups of schools based on the type of science course. The PISA 2006 results showed that girls' in integrated science courses scored slightly higher than scores of girls in subject-specific courses. No differences in boys' scores were found (Åström, 2008). A possible explanation for the small difference in girls' scores offered by Åström was the gender differences in experiences, attitudes, and perceptions of science courses (Åström, 2008).

The earliest study by Åström (2007) of integrated and subject-specific science courses included both qualitative and quantitative components. The quantitative component focused on PISA 2003 science results of 2,000 Swedish students, using HLM. The results showed that the type of science course produced no difference in scores. The qualitative component of the study found the same difference in teacher planning practices as her 2008 study. Teachers of integrated science courses were observed involving students in lesson planning and working with mind maps centered around projects or themes. These practices were not observed in teachers of subject-specific science courses (Åström, 2007).

Faulkner (2012) focused on the levels of integration of middle school science courses and the correlation of the level of integration to students' scores on Connecticut's 8th standardized science assessment. Faulkner (2012) found that, although most Connecticut school districts had adopted an integrated science curriculum for middle schools, most were only partially employing an integrated curriculum. School districts were inconsistent in their implementation of integrated curriculum in terms of its design, planning, implementation, and assessment (Faulkner, 2012). Faulkner also found no significant correlation of level of science course integration and student scores on the state standardized science assessment (2012).

Using four years of data from Massachusetts 8th grade standardized science assessment, Clifford (2016) found no significant difference in student achievement for students taught with an integrated or discipline-based approach. The dissertation by Alwardt (2011) was unique in that it studied the change from a spiral (integrated/comprehensive) science curriculum to a subject-specific science curriculum in Missouri school districts. Although Alwardt did not examine assessment score differences, his study highlights that school district decisions about the type of science curriculum are frequently based on beliefs and opinions rather than on scientific evidence (Alwardt, 2011). Also highlighted were teachers' resistance to change and the need for extensive professional development and support in the successful implementation of a new curriculum (Alwardt, 2011). Table 17 shows the topics and citations reviewed in this section.

Table 17

Science Courses Literature Review Citations:

Topic	Citations	
Comprehensive and integrated science	AAAS, 1989, 2000, 2011.	NAS, 2000.
	Anelli, 2011, 2017.	NGSS Lead States, 2013a.
	Bing Wei, 2009.	NRC, 2012.
	Bjork, 1994.	NRC, 2012.
	Bransford, Brown, & Cocking, 2000.	Pursitasari, Nuryanti, and Rede, 2015.
	Bybee, 1997.	R. E. Mayer, 2008.
	Carr & Harris, 2001.	Ragel, 2015.
	DeBoer, 2000.	Roediger, 2006.
	FLDOE, 2008b, 2015a, 2016e,	Roediger, 2006.
	2017a, 2017c, 2017i, 2017l.	Rutherford, 1990
	G. H. Bower & Hilgard, 1981.	Santau, 2010.
	Halpern & Halpern, 2005.	Silver, Strong, & Perini, 2000.
	Hattie, 2009, 2017.	Stengel, 1997.
	Herr, 2007.	Stengel, 1997.
	IBE-UNESCO, 2017a, 2017b.	Stoica, 2015.
	Kahana, Howard, & Polyn, 2008.	Tamassia and Frans, 2014.
	Kalyuga, Chandler, & Sweller, 1999.	Tyler, 1950.
	Kozma, 2000.	Venville, Wallace, Renne, and Malone, 2002.
	Lovett, 2017.	Zhbanova, Rule, Montgomery, & Nielsen,
	Merritt, 2015.	2010.
Relevant empirical studies	Alwardt, 2011.	
	Åström & Karlsson, 2012.	
	Åström, 2007, 2008.	
	Clifford, 2016.	
	Faulkner, 2012.	
	Tamassia and Frans, 2014.	

Summary

There is no consensus of opinion on which, if either, type of science course leads to higher student achievement and scientific literacy (Tamassia & Frans, 2014). There is some evidence that the type of science course has little effect on student achievement (Åström, 2007, 2008; Åström & Karlsson, 2012; Clifford, 2016; Faulkner, 2012). Due to the dearth of empirical research on this topic, this literature review explored five concepts related to the research questions. The first section, science education, presented the status of science education based on the results of the TIMSS, PISA, NAEP, and FCAT 2.0 Science/SSA, and explored the implications of the results of these assessments from the perspectives of the U.S. STEM workforce pipeline, scientific literacy, and Florida educational leadership. The second section, science assessment, compared the frameworks, standards, and cognitive demands of the PISA, TIMSS, NAEP, and FCAT 2.0 Science/SSA, and looked at the impact of standardized assessment from a student perspective. The third section, science standards, focused on the influence of science standards on scientific literacy. The fourth section, science challenges, presented the challenges of science education for large school districts, low SES students, and ELs. The concluding section, science courses, examined the few empirical studies comparing science achievement of students in comprehensive and subject-specific science courses, based on results of standardized science assessments. Next, Chapter III details the methodology used to determine if there was a difference in student achievement on the 8th grade FCAT 2.0 Science/SSA between two groups of Florida school districts: those that offered comprehensive science courses and those that offered subject-specific science courses, for 2013, 2014, 2015, 2016, and 2017.

CHAPTER III: METHODOLOGY

Introduction

The problem studied was the stagnation of science proficiency of 8^{th} grade science students as measured by required assessments in Florida. The purpose of this quantitative study was to determine if there was a difference in student achievement on the Florida Comprehensive Assessment Test (FCAT) 2.0 Science/Statewide Science Assessment (SSA) from 2013 to 2017 between groups of school districts that offered subject-specific middle grades science courses and groups of school districts that offered comprehensive middle grades science courses. These findings can provide educational leadership with empirical evidence useful for making informed decisions about middle school science course offerings. The independent variable for this study was the middle grades science course type (comprehensive or subject-specific) offered by Florida school districts. The dependent variable was the school district mean scale score for 8^{th} grade students on the FCAT 2.0 Science/SSA for each year from 2013 to 2017. The research questions for this study were:

1. To what extent did 8th grade FCAT 2.0 Science/SSA school district mean scale scores for 2013, 2014, 2015, 2016, and 2017 differ between two groups of Florida school districts: those that offered comprehensive science courses and those that offered subject-specific science courses?

2. To what extent did 8th grade FCAT 2.0 Science/SSA school district mean scale scores for 2013, 2014, 2015, 2016, and 2017 between two groups of Florida school districts: those that offered comprehensive science courses and those that offered subject-specific science courses,

when the school districts are matched by total student population size, low SES student percentage, and EL percentage?

3. To what extent did 8th grade FCAT 2.0 Science/SSA school district mean scale scores for low SES students for 2013, 2014, 2015, 2016, and 2017 differ between two groups of Florida school districts: those that offered comprehensive science courses and those that offered subject-specific science courses, when the school districts are matched by total student population size, low SES student percentage, and EL percentage?

4. To what extent did 8th grade FCAT 2.0 Science/SSA school district mean scale scores for ELs for 2013, 2014, 2015, 2016, and 2017 between two groups of Florida school districts: those that offered comprehensive science courses and those that offered subject-specific science courses, when the school districts are matched by total student population size, low SES student percentage, and EL percentage?

This chapter details the methodology used to test the research questions. The chapter is organized into five sections: (a) selection of subjects; (b) target population; (c) data collection; (d) instrumentation; and (e) test methods, test validity, and data analysis.

Selection of Participants

The participants of this study were selected due to five factors about science education in the state of Florida from 2013 to 2017 that presented an opportunity to help fill the gap in research regarding comprehensive and subject-specific science courses. First, Florida's science education standards and benchmarks remained constant during this period (Florida Department of Education [FLDOE], 2017l). Second, despite a name change from Florida Comprehensive Assessment Test 2.0 Science (FCAT 2.0 Science) to the Statewide Science Assessment (SSA) in the 2015–16 school year, the assessment remained consistent in content, format, administration, complexity level, scale scoring, and achievement level criteria (FLDOE, 2017l). Third, school district FCAT 2.0 Science/SSA scores and demographic data were in the public domain (FLDOE, 2017i). Fourth, course surveys showing the science courses offered each year by each school district were in the public domain (FLDOE, 2013a, 2014b, 2015b, 2016b, 2017f). Fifth, a large population existed of school districts offering each type of science course among the 67 school districts (FLDOE, 2013a, 2014b, 2015b, 2016b, 2017f).

Target Population

The target population was Florida's 8th grade general education students from traditional public (non-charter) middle and junior high schools who took the FCAT 2.0 Science/SSA each year from 2013 through 2017. Students in Florida's four laboratory school districts (Florida A & M University, Florida Atlantic University, and Florida State University); and three special school districts (Florida School for the Deaf and Blind, Florida Virtual School, and Oneida Youth Development Center) were excluded from the target population. Students in charter, virtual, and special schools (correctional facilities, behavioral centers, hospital/homebound, etc.) within the 67 Florida public school districts were also excluded from the target population. While these school districts and schools are required to teach to the Florida NGSSS for science, they are not bound to follow the curricular decisions of the school district in which they reside (§1002.33 (6)(a)2, Fla. Stat., 2016; Florida Consortium of Public Charter Schools, 2017). Low SES student and EL subgroups of the population were also targeted. Table 18 shows the target population numbers of total students assessed, low SES students assessed, and ELs assessed for 2013 through 2017.

Table 18

Target Population and Subgroups

Year	Total students assessed	Low SES students assessed	ELs assessed
2013	165,011	97,246	8,584
2014	166,200	96,399	8,959
2015	166,610	99,849	10,230
2016	157,387	94,155	10,307
2017	154,966	92,001	10,946

Note. Compiled from Florida Department of Education data (FLDOE, 2017i, 2017p).

Data Collection

For each year 2013 through 2017, samples for this study required the 8th grade FCAT 2.0 Science/SSA data categories listed below. These data were gathered and compiled into a master school-district-level database using several publicly-available sources of archival data on the FLDOE website (FLDOE, 2013a, 2014b, 2015b, 2016b, 2017i, 2017p, 2017f, 2017d; FSU, 2017).

1. Year of assessment
2. School-district-level data
3. School district name and number
4. Science course type (comprehensive or subject-specific) offered by traditional public schools within the school district
5. Overall student population

6. Percentage of low SES students
7. Percentage of ELs
8. School-level data
9. School name and number
10. Number of 8^{th} grade students assessed
11. School district mean scale score for all students
12. School district mean scale score for low SES students only
13. School district mean scale score for ELs only

Data Anonymity

Only public record, school-district-level data were used in this study. These data were obtained from Florida's PK-12 Education Information Portal (FLDOE, 2017i). The University of Central Florida's Institutional Review Board determined this study does not constitute human research (see Appendix A). The data contain no individual student identifiers. Florida's PK-12 Education Information Portal and suppress data for schools and school districts with fewer than ten students assessed in any category (FLDOE, 2017i). In this study, a fictitious name was randomly assigned to each school district as an added measure of protection (Fraenkel et al., 2015).

Raw Data Collection

Raw data, in the form of school-level FCAT 2.0 Science/SSA score spreadsheets for each year 2013 through 2017 were downloaded from the Florida Department of Education (FLDOE) FCAT 2.0 Science/SSA School Scores for All Curriculum Groups Grade 8 website (FLDOE, 2017p, 2017i). These data included school-level mean scale scores for every school that administered the 8th grade FCAT 2.0 Science/SSA, including charter and special schools, in Florida's 74 school districts. The individual score spreadsheets for each year were combined into a single spreadsheet of scores for every school in school districts 1 through 74.

Demographic Data Collection, Cleaning, and Compilation

Data were cleaned to ensure only scores for the target population of general education students in traditional public schools were included. All data were removed for schools in school district numbers 68 through 74. These school district numbers are for Florida's four laboratory school districts (Florida A & M University, Florida Atlantic University, Florida State University, and the University of Florida), and three special school districts (Florida School for the Deaf and Blind, Florida Virtual School, and the Oneida Youth Development Center). All data were removed for charter, virtual, and other specialized schools (correctional facilities, behavioral centers, hospital/homebound, etc.) within school district numbers 1 through 67.

The combined spreadsheet was sorted by school name, and compared to a list of Florida charter schools for each year 2013 through 2017, downloaded from the FLDOE website (FLDOE, 2014a). Data for schools that matched the charter school list were removed. To ensure no charter schools remained, the combined spreadsheet was searched for terms such as charter, incorporated, academy, etc. School district websites were checked to verify that the remaining schools were not charter or special schools. Data were removed for any remaining schools outside the target population, such as hospital/homebound, virtual, correctional, behavioral, exceptional, alternative, boys/girls ranch, boys' town, residential, and correctional schools. Data were removed for high schools that administered the 8th Grade FCAT 2.0

Science/SSA retake. School district websites were examined for each high school on the combined spreadsheet to ensure it was a traditional, grades 9 through 12, high school. Scores for these high schools were removed. Some high schools in small school districts include grades 6 through 12. These schools were retained in the combined spreadsheet.

The remaining school-level data were sorted by year, by school district, and compiled into a master school-district-level database. All school-district mean scale scores were calculated using the school-level data of only the traditional public schools. Although it reduced the degrees of freedom for the statistical analyses, these compiled school-district-level data were used for analysis rather than school-level data to maintain consistency with the data limitations for low SES students and ELs.

Neither the school-level score spreadsheets nor the available school-district-level score spreadsheets from the FLDOE website included demographic data or mean scale scores for student subgroups. While school-level data for low SES students and ELs is retrievable from the FLDOE PK-12 EdStats portal, to protect individual student privacy, data are suppressed when the total number of students in a subgroup is fewer than 10 (FLDOE, 2017i). Over 65 percent of schools had FCAT 2.0 Science/SSA data suppressed for ELs for each year. More EL data are available when these suppressed school-level are compiled into school-district-level data.

Due to this limitation, school-district-level FCAT 2.0 Science/SSA mean scale scores for and numbers of low SES students and ELs were retrieved from the FLDOE PK-12 EdStats portal (FLDOE, 2017i). Under “Build Your Own Table,” filters were set to include only 8th grade scores for years 2013 through 2017. The mean scale score data field was selected. Economic and EL status data fields were selected. One-by-one, a filter was set for each school district.

Within each school district, a filter was set for schools, and each traditional public school in each school district was selected. The resulting output tables provided, by year, the school district-level mean scale scores and numbers of students assessed for the low SES student and EL subgroups. These data were added to the master school-district-level database.

Determination of Science Course Type

The science course type offered (comprehensive or subject-specific) by each traditional public school is determined at school-district-level (FLDOE, 2008a). The course type was not included in any of the FCAT 2.0 Science/SSA score data downloaded in previous steps. These data were obtained from course enrollment survey spreadsheets for 2013–2016 downloaded from the FLDOE Data Publications and Reports Archive website, and for 2017 from the FLDOE PK-12 Public School Data Publications and Reports website (FLDOE, 2013a, 2014b, 2015b, 2016b, 2017f). The course names and course numbers for middle school science courses were retrieved from the FLDOE Course Code Directory & Instructional Personnel Assignments website (FLDOE, 2017d).

The course enrollment surveys showed, by year, by school district, student enrollment numbers for every state-approved PK-12 course offered by every school, including charter and special schools. Course enrollment surveys were searched for the middle/junior high school (M/J) science courses shown in Table 19. The data field of course type was added to the master school-district-level database. School districts were numerically coded as either (1) comprehensive or (2) subject-specific, based on the type of science course in which students in each school district were enrolled. The data fields, sources of data, and purposes for the data in the completed master school-district-level database are shown in Table 20. An excerpt from the resulting master school-district-level database is shown in Table 21.

Table 19

Florida Comprehensive and Subject-specific Science Courses

Comprehensive science courses		Subject-specific science courses	
Course number	Course name	Course number	Course name
200204	M/J Compre Sci 1	200101	M/J Erth/Spa Sci
200205	M/J Compre Sci 1 Adv	200102	M/J Erth/Spa Sci Adv
200207	M/J Compre Sci 2	200206	M/J IB MYP Comp Sci1
200208	M/J Compre Sci 2 Adv	200209	M/J IB MYP Comp Sci2
200210	M/J Compre Sci 3	200212	M/J IB MYP Comp Sci3
200211	M/J Compre Sci 3 Adv	200003	M/J IB MYP Life Sci
200205	M/J Compsci1 Acc Hon	200001	M/J Lif Sci
200208	M/J Compsci2 Acc Hon	200002	M/J Lif Sci Adv
		200301	M/J Phy Sci
		200302	M/J Phy Sci Adv
		200002	M/J STEM Life Sci
		200303	M/J STEM Physic Sci

Notes: Compiled from FLDOE 2016-2017 Course Code Directory. Course names shown exactly as listed. IB stands for pre-International Baccalaureate. MYP stands for Middle Years Program (FLDOE, 2017d).

Table 20

Master School District Database Data Fields

Field	Label	Source	Purpose
1	Year of assessment	FLDOE website	Organization
2	School district name	FLDOE website	Unit of analysis
3	Course type	FLDOE Data Publications and Reports Archive website	Independent variable (Comprehensive = 1, Subject-specific = 2)
4	Number of students assessed	FLDOE FCAT website. Sum of students assessed schools included in school district mean scale scores.	Determining target population size and adequacy of sample sizes
5	School district student population	FLDOE EdStats portal. Includes students in all schools within the school district.	Sample matching
6	School district percentage of low SES students	FLDOE EdStats portal. Includes students in all schools within the school district.	Sample matching
7	School district number of low SES students assessed	Estimated by applying school district low SES percentage (7) to number of 8^{th} grade students assessed (5).	Sample matching; calculating low SES population size and adequacy of sample size.
8	School district number of ELs assessed	Estimated by applying school district EL percentage (8) to number of 8^{th} grade students assessed (5).	Sample matching; calculating low SES population size and adequacy of sample size.
9	School district mean scale score	Calculated from school-level mean scale scores from FLDOE website. Includes only data for traditional public schools within school district.	Dependent variable
10	School district low SES mean scale score	Retrieved from EdStats portal. Mean of school-level mean scale scores for only low SES students in traditional public schools within school district.	Dependent variable
11	School district EL mean scale score	EdStats portal. Mean of school mean scale scores for only ELs in traditional public schools within school district.	Dependent variable

Table 21

Master School District 8th Grade FCAT 2.0 Science/SSA Database Excerpt (2013)

School district[a]	Course type	Student population[b]	Low SES percent[c]	Students assessed[c]	Mean scale score[c]	Low SES assessed[c]	Low SES mean scale score[c]	ELs assessed[d]	EL mean scale score[d]
Linwood	Subject-specific	27,826	49.0	1,797	200.4	666	191.1	**	**
Lucas	Comprehensive	4,982	56.8	331	201.0	170	196.4	**	**
Lavender	Comprehensive	26,634	57.2	1,557	196.8	909	180.9	21	167.8
Benson	Comprehensive	3,275	63.3	239	192.0	166	190.1	**	**
Ross	Comprehensive	71,228	45.4	4,334	209.1	1,924	190.8	63	180.7
Giselle	Subject-specific	260,226	56.9	15,953	197.9	8,979	184.5	815	181.8
Madeline	Comprehensive	2,264	66.7	152	205.0	103	200.3	**	**
Samson	Subject-specific	16,355	62.5	1,204	200.5	700	188.9	19	180.7
Rilla	Comprehensive	15,307	63.1	1,092	203.5	615	199.0	**	**
Kimberly	Comprehensive	35,244	36.0	2,792	202.1	1,126	184.9	19	173.1
Barnaby	Subject-specific	43,789	61.2	3,019	200.0	1,894	185.8	189	175.3
Cornell	Subject-specific	9,797	65.1	645	196.0	301	191.9	**	**
Blythe	Comprehensive	4,752	81.6	328	192.0	259	176.2	11	161.7
Zakiah	Comprehensive	2,045	99.4	122	198.0	120	197.8	**	**
Gereon	Comprehensive	125,686	49.1	6,413	195.0	3,936	182.2	187	175.0
Farida	Comprehensive	40,670	61.1	2,707	198.2	1,688	186.4	26	179.9

Note. [a] Fictitious school district names are used. [b] Includes all schools in school district. [c] Includes only traditional public schools in target population. All data compiled from FLDOE Statewide Science Assessment Results (FLDOE, 2017p), FLDOE Course Enrollment Survey (FLDOE, 2013a), and FLDOE EdStats Portal (FLDOE, 2017i). [d] Data masked by asterisks (**) when fewer than 10 students comprise the data (FLDOE, 2017i).

Sampling Method

The sampling method was non-probability, purposive, and stratified (Fraenkel et al., 2015). The sampling unit and unit of analysis were the school district. The school district was chosen as the sampling unit and unit of analysis for four reasons. First, student-level data were not publicly available (FLDOE, 2017i). Second, school-level data were unavailable for many smaller schools with very small populations of low SES students and ELs. Data were suppressed in the FLDOE EdStats system for any score category comprised of ten or fewer students (FLDOE, 2017i). Third, the type of science course used by each school is determined at school-district-level (FLDOE, 2008a). Fourth, the school district unit of analysis facilitated the selection of school district paired samples that offered each type of science course, matched by overall student population, percentage of low SES students, and percentage of ELs.

Sampling was non-probability in that school districts were selected and divided into two groups based upon the type of science course offered. For the Research Questions 2, 3, and 4 paired samples, school districts were matched based on overall student population, percentage of low SES students, and percentage of ELs. All available school district-level data were used for the Research Question 1 sample. Sampling was purposive due to the need for equal numbers of school districts offering each type of science course for the Research Questions 2, 3, and 4 paired samples. Sampling was stratified to select paired samples of school districts that offered different types of science courses, matched by school district student population, percentage of low SES students, and percentage of ELs.

School-district-level data were compiled from the school-level data of only traditional public schools. The compiled school-district-level data were used to stratify school districts into categories by type of science course offered, school district student population, and low SES student percentage. Because the EL percentages of 76 percent of the school districts were clustered at under 10 percent for each year, categorization of the EL percentage demographic was not feasible. Once the school districts were sorted and matched by student population and low SES percentage, school districts with the closest EL percentages were selected as paired samples for Research Questions 2, 3, and 4. Table 22 shows the stratification categories used to select the paired samples.

Table 22

School District Stratification Criteria and Categories

Course type	Total student population categories	Low SES student percentage categories
Comprehensive	Fewer than 12,000	0–49.99
Subject-specific	12,000–41,499	50–59.99
	41,500–99,999	60–69.99
	100,000 and higher	70–100

The criteria used to assign these categories were selected only to facilitate school district paired sample selection for Research Questions 2, 3, and 4. These criteria do not represent interval variables, nor are they related to any published school or school district classification system. The school district size classification system used by the National Center for Education Statistics (NCES) was not appropriate for this study for two reasons. Its most recent categories are based on 2014 data (NCES, 2017d), and it’s twelve-category system does not consider only student population, but assigns size categories based on both the total population of an area and proximity to population centers (NCES, 2006), thus distorting school district size for the purpose of this study. For example, one Florida school district, primarily rural, with a 2016 student population of 8,464 (FLDOE, 2017i), is assigned the NCES locale code for a small city due to its proximity to larger communities in neighboring counties (NCES, 2017d). Two large Florida school districts, one with a 2016 student population of 270,354 (FLDOE, 2017i), and one with a 2016 student population of 358,275 (FLDOE, 2017i), are both assigned the NCES locale code for large suburbs, while a third, with a 2016 student population of just over 100,000 (FLDOE, 2017i) was assigned the NCES locale code for a large city (NCES, 2017d).

Tests performed for Research Question 1 used the 8th grade FCAT 2.0 Science/SSA school district mean scale score data for all 67 Florida school districts for each year's sample. The tests analyzed the differences in the 2013–2017 8th grade FCAT 2.0 Science/SSA school district mean scale scores for all students, low SES students, and ELs, between two groups of Florida school districts: those that offered comprehensive science courses (comprehensive group) and those that offered subject-specific science courses (subject-specific group). Because the number of school districts that offered subject-specific science courses was smaller than the number of school districts that offered comprehensive science courses each year, the Research Question 1 sample groups of comprehensive and subject-specific school districts were unequal.

School District Paired Sample Selection Procedures

While robust to inequalities of variance and sample sizes, independent samples *t*-tests are do not control for other factors that may impact the distributions being tested (Steinberg, 2011). For this reason, Research Questions 2, 3, and 4 used paired samples *t*-tests to control for three demographic factors that have been shown to impact student achievement: school district student population (Amah et al., 2013; Driscoll et al., 2003; Lippman et al., 1996), low socio-economic status (SES) (Becker & Luthar, 2002; Hanushek, 2010; Hattie, 2009; Ladd, 2012; Miller et al., 2013; Wiseman, 2012), English learner (EL) status (Cosentino de Cohen et al., 2005; NCES, 2016). These paired samples required equal numbers of school districts that offered each type of science course, matched by overall student population, low SES student percentage, and EL percentage for each year from 2013 through 2017. Research Question 2 analyzed the differences in FCAT 2.0 Science/SSA mean scale scores for all students in the paired samples. Research Question 3 analyzed FCAT 2.0 Science/SSA school district mean scale scores for low SES students in the paired samples. Research Question 4 analyzed the FCAT 2.0 Science/SSA school district mean scale scores for ELs in the paired samples.

For each year from 2013 to 2017, school district paired samples were purposively selected from the master school-district-level database. The master school-district-level database was sorted by year, then by course type. The master school-district-level database, by year, became the independent samples for Research Question 1. The number of school districts that offered subject-specific science courses each year was less than the number of school district that offered comprehensive science courses. To obtain the largest sample sizes for Research Questions 2 through 4, all subject-specific school districts were selected for the paired samples. Each of these was matched to a comparable comprehensive school district.

Comparability of school districts was triangulated based on three school district demographic criteria: 1) total student population category, 2) low SES student percentage category, and 3) EL percentage. School districts were considered a possible match when they offered different science course types and fell into the same total student population and low SES student percentage categories.

EL percentage was used as a final matching criterion. Of the potential matches based on total student population and low SES student percentage, the school districts with the closest EL percentages were selected. These became the paired samples for Research Questions 2, 3, and 4. If either or both of the paired school districts had suppressed EL mean scale scores, they were removed from the paired sample for Research Question 4. Pearson *r* correlation tests were conducted on each year's paired sample for each matching factor (school district student population, percentage of low SES students, and percentage of ELs) to verify that the school district pairs were well-matched. As shown in Table 23, each year's paired sample had a strong, positive, statistically significant ($\alpha = .01$) correlation (Steinberg, 2011) for each demographic matching factor.

Table 23

School District Paired Samples Demographic Matching Correlations

Year	Demographic	*N* (school district pairs)	*r*	*p*
2013	Student population	18	.95	<.001
	Percentage of low SES students	18	.89	<.001
	Percentage of ELs	18	.88	<.001
2014	Student population	15	.93	<.001
	Percentage of low SES students	15	.90	<.001
	Percentage of ELs	15	.93	<.001
2015	Student population	14	.95	<.001
	Percentage of low SES students	14	.70	<.001
	Percentage of ELs	14	.95	<.001
2016	Student population	15	.98	<.001
	Percentage of low SES students	15	.86	<.001
	Percentage of ELs	15	.73	<.001
2017	Student population	17	.96	<.001
	Percentage of low SES students	17	.70	<.001
	Percentage of ELs	17	.89	<.001

Table 24 shows the Research Questions 2 and 3 paired samples for 2017. Table 25 shows the 2017 Research Question 4 school district paired samples. School district paired samples for 2013 through 2016 for Research Questions 2, 3, and 4 are at Appendix D.

Table 24
Research Questions 2 and 3 School District Paired Sample, 2017 (*N* = 17 school district pairs)

Pair number	School district	Course type[a,]	Student population[b]	Low SES percentage[c]	EL percentage[d]
1	Phokas	Comprehensive	11,542	48.3	1.1
1	Lucas	Subject-specific	5,010	45.6	0.3
2	Firdaus	Comprehensive	5,266	58.7	9.2
2	Viktor	Subject-specific	8,582	49.4	10.5
3	Roel	Comprehensive	9,173	52.7	3.7
3	Cornell	Subject-specific	10,067	57.2	1.3
4	Amias	Comprehensive	2,752	55.3	2.1
4	Lawson	Subject-specific	8,601	58.9	3.0
5	Desta	Comprehensive	5,500	60.3	4.5
5	Katlyn	Subject-specific	6,056	59.9	5.6
6	Blythe	Comprehensive	4,906	64.9	9.1
6	Emmett	Subject-specific	1,268	57.3	7.1
7	Kimberly	Comprehensive	37,052	44.4	2.2
7	Linwood	Subject-specific	29,485	48.5	2.6
8	Lavender	Comprehensive	28,027	47.4	3.1
8	Renato	Subject-specific	31,091	45.2	3.4
9	Gottfried	Comprehensive	12,930	55.1	2.9
9	Samson	Subject-specific	15,925	50.6	2.5
10	Ross	Comprehensive	73,446	50.6	3.6
10	Sulayman	Subject-specific	67,816	47.2	5.0
11	Ciara	Comprehensive	72,490	54.8	4.3
11	Adil	Subject-specific	42,801	47.8	6.7
12	Abioye	Comprehensive	92,682	51.9	9.8
12	Barnaby	Subject-specific	46,407	59.6	15.2
13	Everett	Comprehensive	48,892	54.5	12.9
13	Mirela	Subject-specific	63,023	57.0	19.6
14	Katharine	Comprehensive	40,417	65.2	8.9
14	Placido	Subject-specific	42,516	60.9	4.8
15	Junayd	Comprehensive	63,100	64.4	6.5
15	Walker	Subject-specific	43,040	65.1	6.5
16	Husniya	Comprehensive	214,402	58.6	12.7
16	Giselle	Subject-specific	271,828	61.9	12.7
17	Sunita	Comprehensive	192,729	59.2	12.7
17	Rukiye	Subject-specific	200,667	65.6	14.4

Notes: Fictitious school district names are used. [a]Course Enrollment by School, Survey 3, 2016-17 (FLDOE, 2017f). [b]Student Enrollment by District 2012-17 (FLDOE, 2017e). [c]Economic Status by District 2012-17 (FLDOE, 2017e). [d]ELL Students by District 2012-17 (FLDOE, 2017e)

Table 25
Research Question 4 School District Paired Sample, Year 2017
(N = 11 school district pairs)

Pair number	School district	Course type[a]	Student population[b]	Low SES percentage[c]	EL percentage[d]
1	Firdaus	Comprehensive	5,266	58.7	9.2
1	Viktor	Subject-specific	8,582	49.4	10.5
2	Kimberly	Comprehensive	37,052	44.4	2.2
2	Linwood	Subject-specific	29,485	48.5	2.6
3	Lavender	Comprehensive	28,027	47.4	3.1
3	Renato	Subject-specific	31,091	45.2	3.4
4	Ross	Comprehensive	73,446	50.6	3.6
4	Sulayman	Subject-specific	67,816	47.2	5.0
5	Ciara	Comprehensive	72,490	54.8	4.3
5	Adil	Subject-specific	42,801	47.8	6.7
6	Abioye	Comprehensive	92,682	51.9	9.8
6	Barnaby	Subject-specific	46,407	59.6	15.2
7	Everett	Comprehensive	48,892	54.5	12.9
7	Mirela	Subject-specific	63,023	57.0	19.6
8	Katharine	Comprehensive	40,417	65.2	8.9
8	Placido	Subject-specific	42,516	60.9	4.8
9	Junayd	Comprehensive	63,100	64.4	6.5
9	Walker	Subject-specific	43,040	65.1	6.5
10	Husniya	Comprehensive	214,402	58.6	12.7
10	Giselle	Subject-specific	271,828	61.9	12.7
11	Sunita	Comprehensive	192,729	59.2	12.7
11	Rukiye	Subject-specific	200,667	65.6	14.4

Notes: Fictitious school district names are used. [a]Course Enrollment by School, Survey 3, 2016-17 [b]Student Enrollment by District 2012-17 (FLDOE, 2017e). [c]Economic Status by District 2012-17 (FLDOE, 2017e). [d]ELL Students by District 2012-17 (FLDOE, 2017e).

Sample Sizes

Research Question 1 analyzed the FCAT 2.0 Science/SSA school district mean scale scores for the entire sample (*N* = 67 school districts), by year. Because most Florida school districts offer comprehensive science courses, the Research Question 1 sample groups consisted of unequal numbers of school districts that offered comprehensive science courses (comprehensive group) and school districts that offered subject-specific science courses (subject-specific group). The school district sample group sizes (*n*), by year, for Research Question 1, along with the numbers of 8th grade students assessed represented in the samples, are shown in Table 26.

Table 26

Research Question 1 School District Sample Sizes, by Year

Year	School district group	School districts	Students assessed
2013	Comprehensive	49	108,012
	Subject-specific	18	56,999
		N = 67	165,011
2014	Comprehensive	52	109,891
	Subject-specific	15	56,309
		N = 67	166,200
2015	Comprehensive	53	115,058
	Subject-specific	14	51,552
		N = 67	166,610
2016	Comprehensive	52	107,991
	Subject-specific	15	49,396
		N = 67	157,387
2017	Comprehensive	49*	102,729
	Subject-specific	17	52,237
		N = 66	154,966

Note. *One school district reported no 8th grade FCAT 2.0 Science/SSA results for 2017.

Research Questions 2, 3, and 4 used paired samples of school districts that offered each type of science course, matched by school district student population, percentage of low SES students, and percentage of ELs. Because approximately 75 percent of Florida school districts offered comprehensive science courses, the Research Question 2, 3, and 4 sample sizes (school district pairs) were dependent on the number of school districts that offered subject-specific courses each school year. This reduced the number of school districts available for the Research

Question 2, 3, and 4 paired samples. All school districts offering subject-specific courses were selected for each year's sample. Each of these was paired with a school district that offered comprehensive science courses, matched by school district student population, percentage of low SES students, and percentage of ELs. Research Question 2 analyzed the school district mean scale scores for all students assessed in each year's paired sample. Research Question 3 analyzed the school district mean scale scores for only low SES students. Research Question 4 analyzed the school district mean scale scores for only ELs. Each year's sample size varied as some school districts switched from one type of science course offering to the other. The school district paired sample sizes and numbers of students assessed represented in each year's samples for Research Questions 2 and 3 are shown in Table 27.

Table 27

Research Questions 2 and 3 School District Sample Sizes, by Year

Year	*N* (school district pairs)	Students assessed	Low SES students assessed
2013	18	111,320	64,056
2014	15	111,804	63,220
2015	15	109,183	62,967
2016	15	101,349	59,590
2017	17	108,273	63,018

Research Question 4 analyzed only the mean scale scores for ELs assessed in each school district in the paired samples. Due to data suppression for small school districts with fewer than 10 ELs assessed, not all the school district paired samples could be used. This further reduced the number of school districts available for inclusion in the Research Question 4 paired samples. The school district paired sample sizes and numbers of ELs assessed represented in the samples for Research Question 4 are shown in Table 28.

Table 28

Research Question 4 School District Sample Sizes, by Year

Year	*N* (school district pairs)	ELs assessed
2013	9	5,223
2014	9	5,677
2015	9	6,091
2016	9	6,401
2017	11	7,471

Instrumentation

The instrument used for this study was this study was the Florida Comprehensive Assessment Test 2.0 for Science (FCAT 2.0 Science)/Florida Statewide Science Assessment (SSA). This assessment evaluates science achievement of 5th and 8th grade students every year. The first operational administration of FCAT 2.0 Science took place in 2011 (FLDOE, 2014c). In the 2015-16 school year, FLDOE changed the name from FCAT 2.0 Science to the Statewide Science Assessment (SSA) to make it consistent with the new Florida Standards Assessments (FSA) for mathematics and reading. However, the assessment has remained consistent in format, administration, cognitive complexity, number of items, scoring scale, achievement levels, and benchmarks assessed, since 2011 (FLDOE, 2017j). This study analyzes only 8th grade FCAT 2.0/SSA school district mean scale scores (compiled from school-level mean scale scores for only target population schools) for 2013 through 2017. This section will address the consistency in administration and format; scale score and achievement levels; benchmarks assessed; and accuracy, reliability, and validity of the 8th grade FCAT 2.0 Science/SSA for 2013 through 2017.

Administration and Format

The FCAT 2.0 Science/SSA measures student achievement of Florida's Next Generation Sunshine State Standards for science (NGSSS Science), which were adopted in 2008 (FLDOE, 2017o). Students in grades 5 and 8 participate in the FCAT 2.0 Science/SSA each spring (FLDOE, 2017o). The assessment covers the four content areas of nature of science, physical science, life science, and Earth/space science. The 8th grade FCAT 2.0 Science/SSA is administered as a paper-best test, consisting of 60 to 66 four-option, multiple choice items (FLDOE, 2017e). These include both scored items and non-scored items used for field testing and statistical analysis (FLDOE, 2014c).

From 2013 to 2017, the 8th grade FCAT 2.0 Science/SSA was administered consistently every spring at schools in each school district, within the same range of dates (FLDOE, 2013b, 2014d, 2015d, 2016g, 2017e). Test administrators and proctors were trained and required to sign an agreement to follow the scripted administration manual provided by FLDOE each year dates (FLDOE, 2013b, 2014d, 2015d, 2016g, 2017e). The 8th grade FCAT 2.0 Science/SSA took place in a single day, in two 80-minute sessions, with a stretch break in the middle of each session. Students taking the 8th grade assessment received a hand-held, four-function calculator, and a Periodic Table of the Elements for use during the assessment dates (FLDOE, 2013b, 2014d, 2015d, 2016g, 2017e). By Florida law, schools were required to test at least 95 percent of eligible students (Fla. Admin. Code R. 6A-1.09422, 2016).

Scale Scores and Achievement Levels

The FCAT 2.0 Science/SSA assesses four content categories of nature of science, physical science, life science, and Earth/space science (FLDOE, 2017l). Raw scores (number of correct responses), are used to calculate scale scores, using the statewide 2012 FCAT 2.0 Science scores as the baseline (FLDOE, 2014c). The 120-point scale scores, ranging from 140 to 260 points, are derived using an item response theory (IRT) algorithm (FLDOE, 2014c). This IRT algorithm takes into consideration not only the number of items answered correctly, but also each item's difficulty; cognitive complexity; ease with which the correct answer may be guessed; and ability to differentiate between lower and higher performing examinees (de Ayala, 2008; Kim, 2007; Orr, 2008; Rich, 2017; Visone, 2009, 2010). Because FCAT 2.0 Science/SSA is not an annual assessment, there is no vertical growth component available to gauge a student's progress on the assessment from year-to-year (FLDOE, 2014c). Student achievement on FCAT 2.0 Science/SSA is reported as scale scores and achievement levels, rather than developmental scale scores, as with annual assessments such as mathematics and reading (FLDOE, 2016i).

The scale scoring system and corresponding achievement levels remained constant from 2013 through 2017 (FLDOE, 2013d, 2014h, 2015f, 2016h, 2017e). Achievement levels range from 1 (lowest) to 5 (highest). The minimum passing scale score was 203, which corresponds to Achievement Level 3. The FCAT 2.0 Science/SSA achievement levels, scale scores, and achievement level descriptions are shown in Table 29.

Table 29

FCAT 2.0 Science/SSA Achievement Levels

	Achievement levels				
	1	2	3	4	5
Scale score range	140-184	185-202	203-214	215-224	225-260
Description	Inadequate	Below satisfactory	Satisfactory	Above satisfactory	Mastery

Note. Compiled from annual FLDOE FCAT 2.0 Science/SSA reports (FLDOE, 2013d, 2014h, 2015f, 2016h, 2017e).

In addition to scale scores and achievement levels, raw scores are reported for each content category (nature of science, physical science, life science, and Earth/space science) (FLDOE, 2017h). At school- and school-district-level, these raw scores consist of the average number of items correct in each content category. While the overall scale scoring system remained constant from 2013 through 2017, the number, and cognitive complexity of items measuring standards and benchmarks in each content category varied (FLDOE, 2017h). Because of this, the focus of this study is on school district mean scale scores, compiled from school level mean scale scores for schools that assessed students in the target population, are used in this study. Additional analyses were performed to analyze the raw scores by content category.

Standards and Benchmarks Assessed

Although the content focus of assessment items for each benchmark varied slightly from year to year, the standards and benchmarks assessed in FCAT 2.0 Science/SSA remained consistent from 2013 through 2017 (FLDOE, 2013b, 2014d, 2015d, 2016g, 2017e). For example, each year FCAT 2.0/SSA assesses the following 8th grade nature of science benchmark for standard 1, the practice of science:

> SC.8.N.1.1: Define a problem from the eighth grade curriculum using appropriate reference materials to support scientific understanding; plan and carry out scientific investigations of various types, such as systematic observations or experiments; identify variables; collect and organize data; interpret data in charts, tables, and graphics; analyze information; make predictions; and defend conclusions. (FLDOE, 2012c, p. 38)

In the NGSSS for science numbering scheme, SC designates a science benchmark; 8 designates 8th grade; N designates the nature of science content category; 1 designates the standard—the practice of science; and the final 1 designates the first benchmark for this standard (FLDOE, 2012c). Assessment items for this benchmark may also assess up to seven related 6th, 7th, and 8th grade benchmarks for the same standard (SC.6.N.1.1, SC.6.N.1.3, SC.7.N.1.1, SC.7.N.1.3, SC.7.N.1.4, SC.8.N.1.3, and SC.8.N.1.4). The focus of each year's assessment items for SC.8.N.1.1 may vary among the myriad concepts included in this benchmark or any of its related benchmarks (FLDOE, 2012c).

Assessment items of FCAT 2.0 Science/SSA were organized into four reporting categories used for test design, scoring, and reporting purposes (FLDOE, 2012c). The percentage of raw-score points derived from each content category, specified in each year's test design summary, remained constant from 2012-17 (FLDOE, 2017l). For the for the 8th grade FCAT 2.0 Science/SSA, this percentage was 19 percent nature of science, 27 percent physical science, 27 percent life science, and 27 percent Earth/space science (FLDOE, 2013b, 2014d, 2015d, 2016g, 2017e).

Florida's NGSSS for science are comprised of 18 standards. The standards are broken down by content category into 109 benchmarks (FLDOE, 2012c). Of the total 109 middle grades science benchmarks, 15 were designated as "not assessed" in FCAT 2.0 Science/SSA (FLDOE, 2012c, p. B-1). Of the 94 assessed benchmarks, 90 were assessed annually, either individually or in combination with other benchmarks belonging to the same standard (FLDOE, 2012c). Only four benchmarks were assessed intermittently (FLDOE, 2012c). The numbers of benchmarks assessed and not assessed in the 8th grade FCAT 2.0 Science/SSA remained consistent from 2013 through 2017, and are shown, by content category, in Table 30.

Table 30

8th Grade FCAT 2.0 Science/SSA, Benchmarks Assessed/Not Assessed

	Number of benchmarks		
Content category	Assessed	Not assessed	Total
Nature of science	23	11	34
Physical science	24	0	24
Life science	20	1	21
Earth/space science	27	3	30
Total	94	15	109

Note. Data compiled from publicly-available FCAT 2.0 Science Test Item Specifications, Version 2, Grade 8 (FLDOE, 2012c), and CPALMS (FSU, 2017).

Items on the 2013 through 2017 FCAT 2.0 Science/SSA were categorized also by cognitive complexity, which refers to the cognitive demand associated with an item (FLDOE, 2013b, 2014d, 2015d, 2016d, 2017l). Low-complexity items involved recall and recognition.

Moderate-complexity items involved flexible thinking, reasoning, and problem-solving skills. High-complexity items involved analysis and abstract reasoning (FLDOE, 2013b, 2014d, 2015d, 2016d, 2017l). The 2013 through 2017 8th grade FCAT 2.0 Science/SSA derived 10 to 20 percent of points from low-complexity items, 60 to 80 percent from moderate-complexity items, and 10 to 20 percent from high-complexity items (FLDOE, 2013b, 2014d, 2015d, 2016d, 2017l).

Reliability

Reliability is the consistency of results obtained from an assessment (Lunenburg & Irby, 2008). Both reliability and validity (addressed in the next section) are essential for making appropriate interpretations of FCAT 2.0/SSA scores (FLDOE, 2016d). Measurements of the reliability of FCAT 2.0 Science for 2013 and 2014 were published in the Florida Statewide Assessments 2014 Technical Report (FLDOE, 2014c). Measurements of the reliability of FCAT 2.0 Science and SSA for 2014, 2015, and 2016 were published in the Florida Statewide Science and EOC Assessments 2016 Technical Report (FLDOE, 2016d). The 2017 SSA technical report will be published in early 2018 (Cirio, 2017).

Reliability of an assessment can be measured using several methods, all involving measuring the consistency of scores for the same students (Fraenkel et al., 2015). The test-retest method administers the same assessment twice to the same students. If each student's scores are consistent on both, the assessment is considered reliable (Fraenkel et al., 2015). The equivalent forms method administers two different but equivalent versions of an assessment to the same students. If each student's scores are consistent on both versions, the assessment is considered reliable (Fraenkel et al., 2015). Due to time and resource constraints, neither test-retest nor equivalent forms reliability could be used for FCAT 2.0 Science/SSA (FLDOE, 2014c, 2016d).

Instead, the reliability of the 8th grade FCAT 2.0 Science/SSA was estimated with two measurements of the internal consistency of students' scores on a single assessment (FLDOE, 2014c, 2016d). Cronbach's alpha was used to measure the consistency of scores among items assessing similar content. Marginal reliability was used to measure the consistency of students' scores among items of the same cognitive complexity level (FLDOE, 2014c, 2016d). Both of these estimates of reliability require at least 40 items on the assessment (Lunenburg & Irby, 2008). The 8th grade FCAT 2.0 Science/SSA contains 60 to 66 items. These estimates of reliability also require unidimensional assessment items, meaning that the items measure the same thing (Hattie, 1985). Evidence of the unidimensionality of the 8th grade FCAT 2.0 Science/SSA is strong and is addressed later in the validity section of this chapter.

Cronbach's alpha is a measure of the average internal consistency of each student's scores for groups of similar items on an assessment (Fraenkel et al., 2015). For example, if all questions assessing nature of science benchmarks yielded very similar scores for each student, the internal consistency of that group of questions would be high. If all groups of questions on an assessment have high internal consistency, the overall consistency, or Cronbach's alpha, for the assessment would be high. Best and Kahn (1998) offer a criterion for evaluating Cronbach's alpha, which ranges from 0 to 1. A Crohnbach's alpha of .00 to .20 is "negligible"; .20 to .40 is "low"; .40 to .60 is "moderate"; .60 to .80 is "substantial"; and above .80 is "high to very high" (Best & Kahn, 1998, p. 372). Lunenburg and Irby (2008) state that a Cronbach's alpha of 0.8 or higher is acceptable for assessment with 40 or more items. Using these criteria, the Cronbach's alpha of the 8th grade FCAT 2.0 Science/SSA for each year from 2013 through 2016 was very high, as shown in Table 31.

Marginal reliability measures the consistency of questions of the same cognitive complexity levels (de Ayala, 2008). If all questions of the same cognitive complexity level on an assessment yield consistent scores for each student, the marginal reliability of the assessment is high. As with Cronbach's alpha, measures of marginal reliability range from 0 to 1 (Best & Kahn, 1998). Using Best and Kahn's (1998) criterion for evaluating Cronbach's alpha for evaluating marginal reliability, Table 31 shows the marginal reliability of the 8th grade FCAT 2.0 Science/SSA for each year from 2013 through 2016 was very high.

Table 31

8th Grade FCAT 2.0 Science/SSA Reliability, 2013-2016

	2013	2014	2015	2016
Cronbach's alpha	.913	.895	.905	.910
Marginal reliability	.941	.942	.940	.940

Note. Data compiled from FLDOE FCAT 2.0 Science/SSA Technical Reports and Yearbooks (FLDOE, 2013b, 2014d, 2015d, 2016d).

The 8th grade FCAT 2.0 Science/SSA technical reports also provide Cronbach's alpha coefficients for the following subgroups: female, male, gender unknown, African-American, American Indian or Alaska Native, Asian, Hispanic/Latino, multi-racial, Native Hawaiian or Pacific Islander, White, ethnicity unknown, and English learners. The lowest Cronbach's alpha for all these groups, from 2013 through 2016, was .862 for English learners. This was high according to the criteria of both Best and Kahn (1998) and Lunenburg and Irby (2008).

Validity

Validity is the degree to which an assessment measures what is intended to be measured (Lunenburg & Irby, 2008). Validation of an assessment instrument involves the collection of multiple forms evidence supporting the validity of the instrument (Onwuegbuzie, 2000). The more evidence gathered, the more valid the instrument is deemed to be (Fraenkel et al., 2015). Evidence of the validity of FCAT 2.0 Science for 2013 through 2016 was published in the Florida statewide assessments technical reports and yearbooks (FLDOE, 2013b, 2014c, 2014d, 2015d, 2016d). The 2017 SSA technical reports will be published in early 2018 (Cirio, 2017). Although the reliability and validity analysis for the 2017 SSA has not yet been published, there were no significant changes content, test item specifications, administration, format, or administration of 8th grade FCAT 2.0 Science/SSA from previous years (FLDOE, 2017e).

There are three main types of evidence of validity: content-related, criterion-related, and construct-related (Fraenkel et al., 2015). Content-related evidence of validity, or face validity, refers to the content, format, and comprehensiveness of an assessment (Lunenburg & Irby, 2008). Content-related evidence shows that the assessment items measure the intended standards and benchmarks; the items are presented in a manner consistent with specified wording, format, and cognitive complexity; and measure all of the intended standards and benchmarks (FLDOE, 2014c; Fraenkel et al., 2015). Criterion-related evidence shows that scores for an assessment are highly correlated with other validated assessments of the same content (Lunenburg & Irby, 2008). Construct-related evidence refers to the extent to which an assessment item measures an underlying characteristic in the students taking the assessment (FLDOE, 2014c; Lunenburg &

Irby, 2008). The technical reports for the FCAT 2.0 Science/SSA address all three types of evidence of validity.

Content-related Evidence of Validity

Content-related evidence of validity is presented in the technical reports for each year's assessment (FLDOE, 2014c, 2016d). These reports, approximately 150 pages long, document every aspect of FCAT 2.0 Science/SSA item development. Included in the technical reports are chapters regarding: the participants in the development and analysis of the assessments; item development, review, and field testing processes; assessment administration; interpretation and limitations of assessment reports; performance standards; scale scoring; reliability, validity, and quality control (FLDOE, 2014c, 2016d). Almost every page of the technical reports supports that the assessment items measure the intended standards and benchmarks (FLDOE, 2014c, 2016d).

Items on the 8th grade FCAT 2.0 Science/SSA were based on NGSSS for science content standards and benchmarks, along with detailed test item specifications that define what is and is not to be assessed (FLDOE, 2014c, 2016d). Each year, assessment items were developed by teams of subject-matter experts, test development experts, and the FLDOE Test Development Center (TDC) following a comprehensive item development plan (IDP) (FLDOE, 2014c, 2016d). Item development team members were trained by Pearson Education, Inc. Test Development and Psychometric Services and the FLDOE TDC (FLDOE, 2014c, 2016d; Pearson VUE, 2014). The IDP specified the number of items to be developed by subject, grade, content category, benchmark, and cognitive complexity level (FLDOE, 2014c, 2016d). Every item went through an 18 to 24 month validation process conducted by content experts, psychometricians, and

Florida educators (FLDOE, 2014c, 2016d). Teams of Florida educators representing each of the state's geographic regions and cultural groups reviewed every item for potential threats to validity, such as biases regarding gender, disability, racial, ethnic, linguistic, religious, geographic, and socio-economic groups (FLDOE, 2012c). Items were again reviewed for content accuracy and currency by an independent third party at Florida State University's Center for Advancement of Learning and Assessment (FLDOE, 2014c, 2016d).

The reliability estimates (Cronbach's alpha and marginal reliability, addressed in the reliability section of this chapter) show that the scale scoring system of the 8th grade FCAT 2.0 Science/SSA was highly consistent from 2013 through 2016. Student scores were consistent among groups of items measuring the same content standards, and among groups of items of the same cognitive complexity levels. This reliability bolsters the 8th grade FCAT 2.0 Science/SSA content-related evidence of validity (FLDOE, 2014c, 2016d).

Criterion-related Evidence of Validity

The FCAT 2.0 Science/SSA technical reports state that lack of criterion-related evidence of validity is a weakness, and address two obstacles to gathering this type of evidence (FLDOE, 2014c, 2016d). First, because the 8th grade FCAT 2.0 Science/SSA was designed to assess student achievement of Florida's NGSSS for science, finding a fully valid criterion based on these standards is difficult (FLDOE, 2014c, 2016d). Correlation of FCAT 2.0 Science/SSA with some portions of other assessments, such as ACT, SAT, or another state's science assessment, may be possible. However, the standards for those assessments differ from Florida's NGSSS for science. Because of this, the criterion-related evidence would not be strong (FLDOE, 2014c, 2016d). Second, the only way to find a fully-valid criterion would be to develop and validate an entire second assessment (FLDOE, 2014c, 2016d). Because of these two obstacles, no criterion-related evidence was presented in the 8th grade FCAT 2.0 Science/SSA technical reports or yearbooks. The reports do state that the TDC continues to look at external sources, such as school districts conducting program evaluation research, university researchers, and special interest groups, to collect criterion-related evidence of validity (FLDOE, 2014c, 2016d). The lack of criterion-related validity limits this in its usefulness to school districts outside the state of Florida.

Construct-related Evidence of Validity

Construct-related validity, which measures an assessment's ability to gauge an underlying student characteristic, requires a wide variety of evidence, and "the more and the more varied, the better" (Fraenkel et al., 2015, p. 154). The general construct intended to be measured by the 8th grade FCAT 2.0 Science/SSA is student achievement and knowledge of 6th, 7th, and 8th grade NGSSS for science (FLDOE, 2017c). As construct-related evidence, the FCAT 2.0 Science/SSA technical reports used several different models to validate the unidimensionality and the scoring of the items (FLDOE, 2014c, 2016d).

Unidimensionality of assessment items means that all items, or groups of items, on an assessment measure the same construct (Hattie, 1985). The items on FCAT 2.0 Science/SSA measured the construct of student achievement and knowledge of middle grades standards and benchmarks in four content categories of science: nature of science, physical science, life science, and earth/space science (FLDOE, 2017c). Items are grouped into these categories and scores were measured for consistency using three confirmatory factor analysis (CFA) models: 1) root mean square error of approximation (RMSEA); 2) comparative fit index (CFI); and 3)

Tucker-Lewis index (TLI) (FLDOE, 2014c, 2016d). Each of these CFA models analyzes the consistency, or fit, of student scores among items in the same content categories (Rigdon, 1996). Resulting indices of all three models may range from 0 to 1. For the RMSEA model, indices close to 0 mean that there is very little misfit among groups of items measuring the same construct (FLDOE, 2014c, 2016d; Rigdon, 1996). For the CFI and TLI models, indices close to 1 indicate a high level of consistency among groups of items measuring the same construct (T. A. Brown, 2006; FLDOE, 2014c, 2016d; Rigdon, 1996; L. R. Tucker & Lewis, 1973). The three CFA indices (RMSEA, CFI, and TLI) shown in Table 32 indicate high construct-related validity for the 8th grade FCAT 2.0 Science/SSA from 2013 through 2016.

Table 32

CFA Model Fit Summary, FCAT 2.0 Science/SSA, 2013-16

Year	RMSEA	CFI	TLI
2013	.015	.986	.986
2014	.015	.986	.986
2015	.015	.986	.986
2016	.015	.986	.986

Note. Data compiled from FCAT 2.0 Science/SSA technical reports (FLDOE, 2014c, 2016d). Indices for 2017 to be published by FLDOE in January 2018 (Cirio, 2017).

Test Methods, Test Validity, and Data Analysis

Samples of school district mean scale scores were analyzed using *t*-tests to answer four research questions. The *t*-test was chosen due to its robustness to extremely small sample sizes, inequalities of variance, and inequalities of sample sizes (De Winter, 2013; Kohr & Games, 1974; Steinberg, 2011). Factorial ANOVA was not used because two of the three control factors selected (school district student population school district EL percentage) were not normally distributed among the 67 school districts. This prevented the formation of roughly equal groups required for factorial ANOVA (Steinberg, 2011).

The independent variable for all four research questions was science course type (comprehensive of subject-specific). The dependent variable for all four research questions was school district mean scale score. Tests were conducted on the independent variable, by year, for three groups of students: all students, low SES students, and ELs. All tests were non-directional, $\alpha = .05$. All tests were conducted such that M_1 was the school district mean scale score for school districts that offered comprehensive science courses, and M_2 was the school district mean scale score for school districts that offered subject-specific science courses.

Research Question 1 used independent samples *t*-tests (Steinberg, 2011) to analyze the difference in school district mean scale scores for all students, low SES students, and ELs, between two groups of school districts: those that offered comprehensive science courses and those that offered subject-specific science courses. No controls were used for school district population, percentage of low SES students, percentage of ELs, or any other demographic factor. Because the number of school districts that offered subject-specific science courses was smaller than the number of school districts that offered comprehensive science courses each year, sample sizes were unequal.

Although independent samples *t*-tests are robust to unequal sample sizes, extremely small sample sizes, and inequalities of variance in the samples (De Winter, 2013; Kohr & Games, 1974; Steinberg, 2011), sample data were analyzed for normality of distribution (skewness and kurtosis) and equality of variances (using Levene's test) to support the validity of the independent samples *t*-tests (Steinberg, 2011). When Levene's test did not confirm the equality of variance in the samples, the degrees of freedom used in the independent samples *t*-tests were reduced (Kohr & Games, 1974; Steinberg, 2011). When only two groups are being analyzed, the Statistical Package for the Social Sciences (SPSS®) independent samples *t*-test for unequal variances provides the same results as Welch's *t*-test (IBM Support, 2016). Welch's *t*-test is very robust to unequal sample sizes and variances because it weights the variances of the two samples proportionally to the number of subjects in each sample (Kohr & Games, 1974; Steinberg, 2011). For statistically significant results, effect sizes were calculated using Cohen's *d* (Steinberg, 2011).

Independent samples *t*-tests are not designed to control for any other factors that may impact the distributions being tested (De Winter, 2013; Steinberg, 2011). For this reason, Research Questions 2, 3, and 4 used paired samples *t*-tests to control for three demographic factors that have been shown to impact student achievement: school district student population (Amah et al., 2013; Driscoll et al., 2003; Lippman et al., 1996), low socio-economic status (SES) (Becker & Luthar, 2002; Hanushek, 2010; Hattie, 2009; Ladd, 2012; Miller et al., 2013; Wiseman, 2012), and English learner (EL) status (Cosentino de Cohen et al., 2005; NCES, 2016). Paired samples *t*-tests (Steinberg, 2011) were used to analyze school district mean scale score differences between paired samples of Florida school districts that offered comprehensive science courses to Florida school districts that offered subject-specific science courses. Each school district pair consisted of a school district that offered comprehensive science courses and a school district that offered subject-specific science courses, matched by student population, percentage of low SES students, and percentage of ELs.

Research Question 2 analyzed the difference in mean scale scores for all students in between the school district paired samples. Research Question 3 analyzed the mean scale scores for only low SES students between the school district paired samples. Research Question 4 analyzed the difference in mean scale scores for only ELs between the school district paired samples. All data were analyzed for normality of distribution (skewness and kurtosis) and homogeneity of variance (using Pitman-Morgan tests) to support the validity of the paired samples *t*-tests (Gardner, 2001; Kohr & Games, 1974; Morgan, 1939; Pitman, 1939; Steinberg, 2011). Effect sizes for statistically significant *t*-test results were calculated using Cohen's *d* (Steinberg, 2011). The research questions, variables, and test methods used in this study are summarized in Table 33.

Table 33

Summary of Research Questions, Target Population, Sample Sizes, Variables, and Tests

Research question	Target population	Sample size	Independent variable	Dependent variable	Controls	Test
1. To what extent did school district mean scale scores differ between independent samples of the two school district groups?	All 8th grade students assessed	School districts 2013: $N = 67$ 2014: $N = 67$ 2015: $N = 67$ 2016: $N = 67$ 2017: $N = 66$	Science course type (comprehensive subject-specific)	School district FCAT 2.0 Science/SSA mean scale score for all 8th grade students	None	Independent samples *t*-test; Levene's test; Skewness/kurtosis analysis; Cohen's *d*
2. To what extent did school district mean scale scores for all students differ between paired samples of the two school district groups?	All 8th grade students assessed	School district pairs 2013: $N = 18$ 2014: $N = 15$ 2015: $N = 15$ 2016: $N = 15$ 2017: $N = 17$	Science course type (comprehensive or subject-specific)	School district FCAT 2.0 Science/SSA mean scale score for all students	School district student population, low SES percentage, EL percentage	Paired samples *t*-test; Pitman-Morgan test; Skewness/ kurtosis analysis; Cohen's *d*
3. To what extent did school district mean scale scores for low SES students differ between paired samples of the two school district groups?	All 8th grade low SES students assessed	School district pairs 2013: $N = 18$ 2014: $N = 15$ 2015: $N = 14$ 2016: $N = 15$ 2017: $N = 17$	Science course type (comprehensive or subject-specific)	School district FCAT 2.0 Science/SSA mean scale score for low SES students	School district student population, low SES percentage, EL percentage	Paired samples *t*-test; Pitman-Morgan test; Skewness/ kurtosis analysis; Cohen's *d*
4. To what extent did school district mean scale scores for ELs differ between paired samples of the two school district groups?	All 8th grade ELs assessed	School district pairs 2013: $N = 9$ 2014: $N = 9$ 2015: $N = 9$ 2016: $N = 9$ 2017: $N = 11$	Science course type (comprehensive or subject-specific)	School district FCAT 2.0 Science/SSA mean scale score for ELs	School district student population, low SES percentage, EL percentage	Paired samples *t*-test; Pitman-Morgan test; Skewness/ kurtosis analysis; Cohen's *d*

Summary

This chapter detailed the selection of participants, data collection, instrumentation, test methods and data analysis employed in this study. The problem studied was the stagnation of science proficiency of 8th grade science students as measured by required assessments in Florida. The purpose of this quantitative study was to determine if there was a difference in student achievement on the 8th grade FCAT 2.0 Science/SSA between two groups of school districts: those that offered comprehensive middle grades science courses and groups of school districts that offered subject-specific middle grades science courses from 2013 to 2017.

The target population was Florida's 8th grade general education students from traditional public (non-charter) middle and junior high schools who took the FCAT 2.0 Science/SSA each year from 2013 through 2017. Two subgroups of this population were also targeted: low SES students and ELs. The 8th grade FCAT 2.0 Science/SSA score data for 2013 through 2017 were collected from publicly-available sources on the FLDOE website. School district mean scale scores were compiled from school-level data of only schools with students in the target populations. Data for charter schools and other special schools were removed.

The sampling method was non-probability, purposive, and stratified. Research Question 1 used all available 8th grade FCAT 2.0 Science/SSA school district mean scale score data for the 67 Florida school districts as the sample. Research Questions 2, 3, and 4 used paired samples of school districts that offered comprehensive science courses and school districts that offered subject-specific science courses, matched by student population, percentage of low SES students, and percentage of ELs.

Independent samples *t*-tests were used for Research Question 1 to analyze differences in 8th grade FCAT 2.0 Science/SSA mean scale scores for all students, low SES students, and ELs between groups of Florida school districts that offered comprehensive science courses and groups that offered subject-specific science courses for each year from 2013 through 2017. Research Questions 2, 3, and 4 used paired samples *t*-tests to analyze differences in 8th grade FCAT 2.0 Science/SSA school district mean scale scores between paired samples of school districts that offered comprehensive science courses and those that offered subject-specific science courses, matched by student population, percentage of low SES students, and percentage of ELs. Research Question 2 analyzed the school district mean scale scores for all students in the paired samples. Research Question 3 analyzed school district mean scale scores for only low SES students in the paired samples. Research Question 4 analyzed the school district mean scale scores for only ELs in the paired samples.

The instrumentation used for this research was the 8th grade FCAT 2.0 Science/SSA for each year from 2013 through 2017. Except for a name change after 2015, this instrument remained consistent from 2013 through 2017. The 8th grade FCAT 2.0 Science/SSA technical reports for each year documented evidence of reliability, content validity, and construct validity. Threats to internal validity were addressed, and these threats were minimal. While the test methods used are robust to unequal sample sizes and unequal variances (Kohr & Games, 1974; Steinberg, 2011), samples were analyzed to verify distribution normality and equality of variance. The results of the tests detailed in this chapter are presented next in Chapter IV.

CHAPTER IV: FINDINGS

Introduction

The problem studied was the stagnation of science proficiency of 8th grade science students as measured by required assessments in Florida. The purpose of this study was to determine if there was a difference in student achievement on the Florida Comprehensive Assessment Test 2.0 for Science (FCAT 2.0 Science)/Statewide Science Assessment (SSA) for each spring assessment from 2013 to 2017 between two groups of school districts: those that offered comprehensive middle grades science courses and those that offered subject-specific middle grades science courses. The dependent variable for this study was the 8th grade FCAT 2.0 Science/SSA school district mean scale score for each year from 2013 through 2017. The independent variable was the middle grades science course type (comprehensive or subject-specific) offered by Florida school districts.

This chapter presents the results of tests for statistical significance in the differences for the 2013, 2014, 2015, 2016, and 2017 8th grade FCAT 2.0 Science/SSA school district mean scale scores for three groups of students: all students, low SES students, and ELs; between two groups of school districts: those that offered comprehensive science courses (comprehensive group), and those that offered subject-specific science courses (subject-specific group). All tests were non-directional, $\alpha = .05$. The p values were rounded to two digits, except for cases where the results approach statistical significance. In these cases, p values were rounded to three digits. All tests were conducted such that M_1 was the school district mean scale score for the comprehensive group, and M_2 was the school district mean scale score for the subject-specific group. Thus, a positive *t* value indicated that the mean scale score for the comprehensive group was numerically higher than that of the subject-specific group, and vice versa.

This chapter presents the results of tests for statistical significance in the differences in the 2013, 2014, 2015, 2016, and 2017 8th grade FCAT 2.0 Science/SSA school district mean scale scores for the three groups, between the two groups of school districts. All tests were non-directional, $\alpha = .05$. The *p* values were rounded to two digits, except for cases where the results approached statistical significance. In these cases, *p* values were rounded to three digits. All tests were conducted such that M_1 was the school district mean scale score for the comprehensive group, and M_2 was the school district mean scale score for the subject-specific group. Thus, a positive *t* value indicated that the mean scale score for the comprehensive group was numerically higher than that of the subject-specific group, and vice versa.

Research Question 1 used all available 8th grade FCAT 2.0 Science/SSA mean scale score data for all Florida school districts (excluding the special school districts). Independent samples *t*-tests were used to test for statistical significance in the difference in school district mean scale scores for the three student groups between the two groups of school districts.

Research Questions 2, 3, and 4 used paired samples of school districts, one from each group of school districts, matched by student population, percentage of low SES students, and percentage of ELs. Research Question 2 tested the differences in school district mean scale scores for all students between the two school district groups in the paired samples. Research Question 3 tested the differences in school district mean scale scores for low SES students between the two school district groups in the paired samples. Research Question 4 tested the differences in school district mean scale scores for ELs between the two school district groups.

For each research question, descriptive statistics are presented, followed by findings and analyses for the tests conducted for each year's sample.

This chapter begins with a look at the descriptive statistics of the Florida school districts from which the research question samples were drawn. Next, the research question test results are presented, organized by research question, by student group, by year. Finally, results of additional analyses are presented.

Descriptive Statistics

The descriptive statistics of Florida's 67 school districts spanning the five years from 2013 to 2017, of Florida's 67 school districts are shown in Table 34, both overall and by school district group. During this period, an average of 51 school districts (76 percent) offered comprehensive courses, and 16 (24 percent) offered subject-specific courses. The mean student populations, percentages of low SES students, and percentages of ELs were comparable between the two school district groups. The school district mean scale scores for all three student groups were numerically higher in the subject-specific group, but less than the minimum passing score of 203 in both school district groups. The mean pass rate for all students was numerically higher in the subject-specific group, but less than 50 percent in both school district groups.

Table 34

Florida School District Descriptive Statistics, 2013–2017 (*N* = 67 school districts)

		Student population			Mean scale score			
School district group	Statistic	Total	Low SES %	EL %	All students	Low SES students	ELs	Pass rate
Comprehensive (*n* = 51 school districts)	*M*	37,444	59.73	5.43	198.09	191.20	178.92	44.18
	Mdn	11,161	59.10	3.94	199.80	190.54	178.09	46.87
	SD	64,993	12.32	4.75	8.38	6.17	5.88	11.77
	Min	889	23.19	0.63	155.80	182.43	161.73	17.40
	Max	356,659	99.40	20.72	209.38	202.38	191.85	66.20
Subject-specific (*n* = 16 school districts)	*M*	51,194	56.09	6.49	200.67	190.20	179.40	48.85
	Mdn	30,351	59.71	4.79	200.31	188.47	179.88	47.44
	SD	71,120	6.90	5.17	2.93	4.22	3.22	6.34
	Min	1,238	43.08	0.26	196.83	185.08	173.82	40.20
	Max	266,365	65.87	18.42	206.19	197.20	185.59	59.79
Total	*M*	40,933	58.81	5.73	198.74	190.95	179.06	45.37
	Mdn	12,855	59.23	4.30	199.80	189.49	178.70	46.94
	SD	66,322	11.26	4.85	7.45	5.72	5.23	10.80
	Min	889	23.19	0.26	155.80	182.43	161.73	17.40
	Max	356,659	99.40	20.72	209.38	202.38	191.85	66.20

As shown in Figure 3, the overall distribution of student population was positively skewed among the 67 school districts. Most school districts had student populations of fewer than 75,000 students. Half had student populations of fewer than 13,000 students.

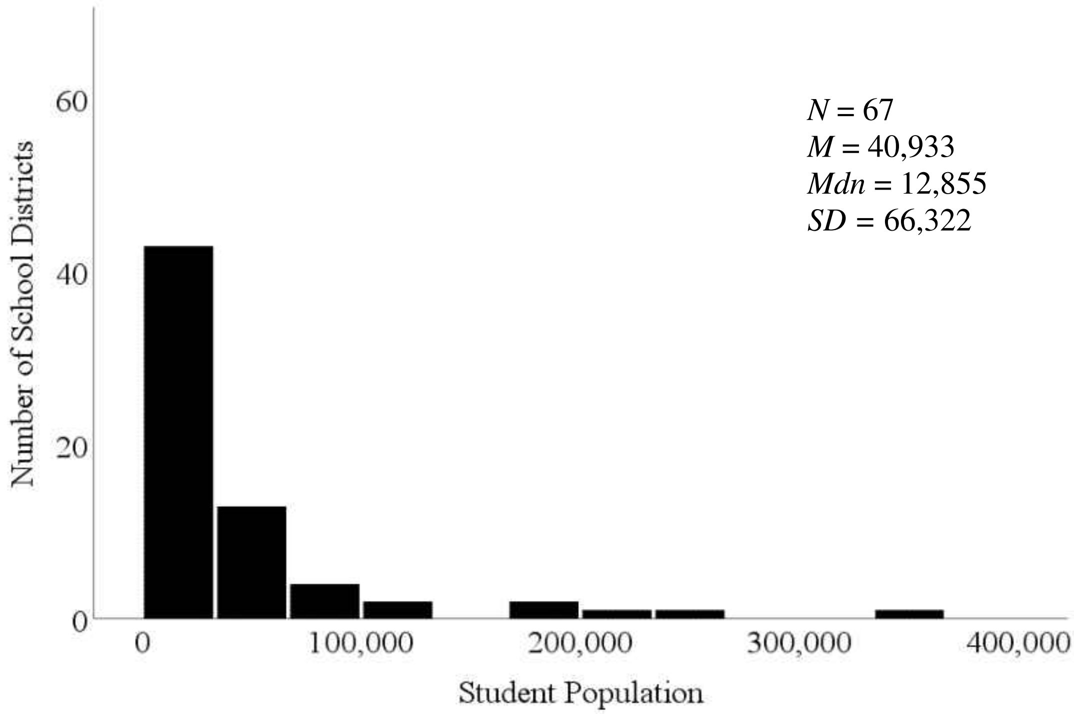

Figure 3. Mean student population of Florida's 67 school districts, 2013–2017.

As shown in Figure 4, the distribution of low SES students, as a percentage of total student population, was relatively normal among the 67 school districts. The mean and median were both at about 60 percent. Low SES students comprised 40 to 80 percent of the student populations of most Florida school districts.

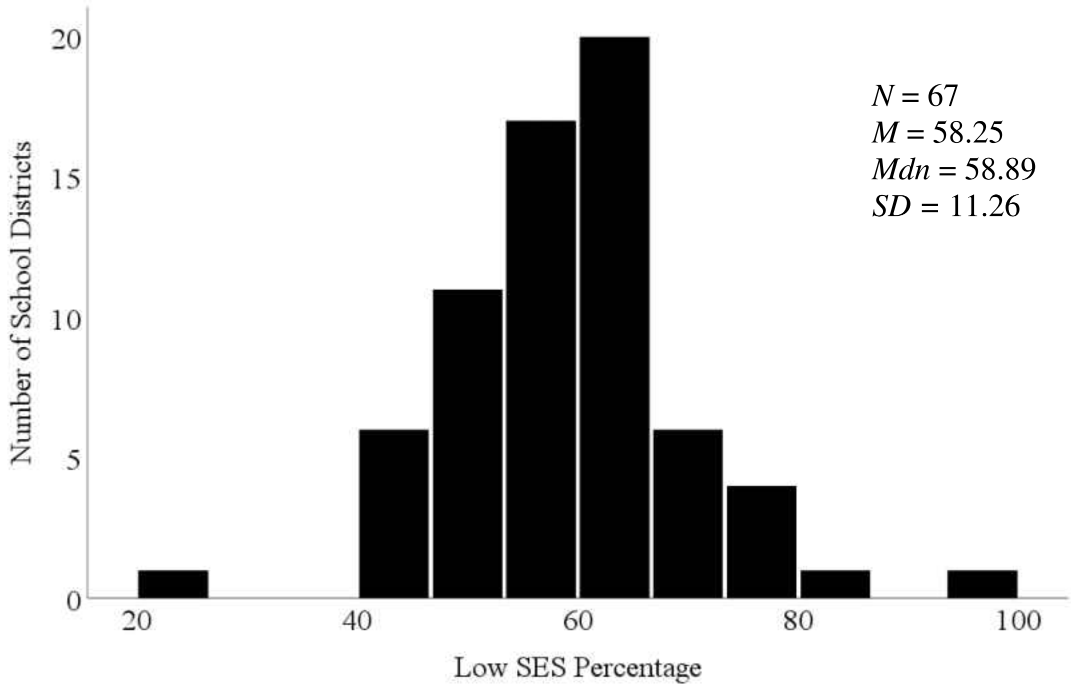

Figure 4. Mean low SES student percentage of Florida's 67 school districts, 2013–2017.

As shown in Figure 5, the distribution of ELs as a percentage of total student population was positively skewed. English learners comprised less than ten percent of the student populations of most school districts. In half of the school districts, ELs comprised less than 5 percent of the student population.

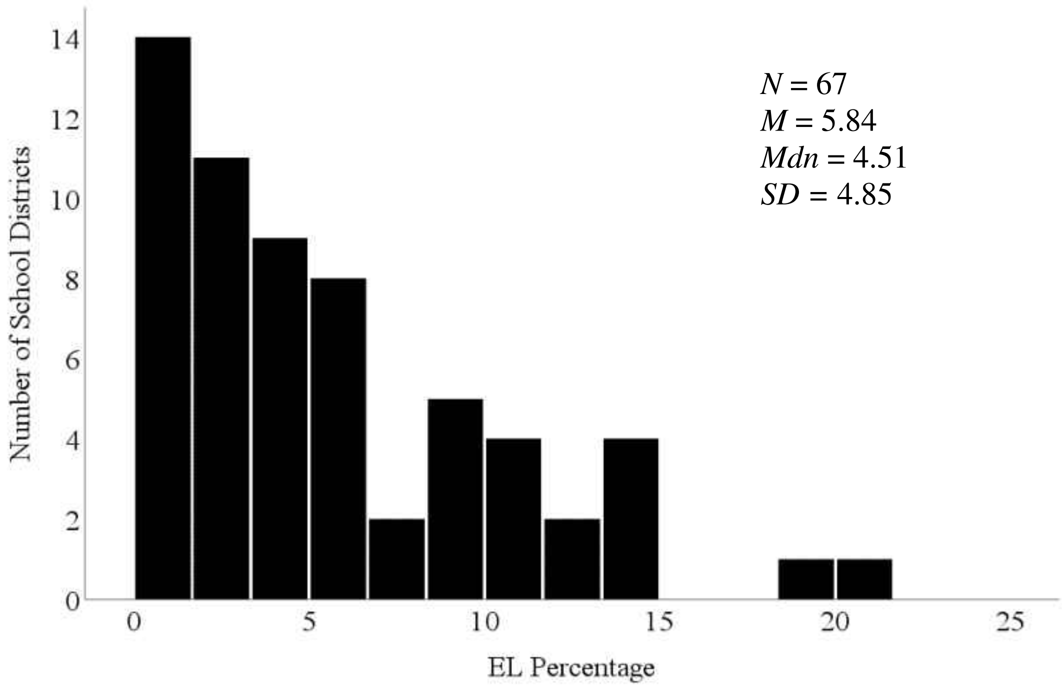

Figure 5. Mean EL percentage of Florida's 67 school districts, 2013–2017.

From 2013 to 2017, only 17 of the 67 school districts had 2013–2017 mean scale scores for all students that reached or exceeded the minimum passing score. Descriptive statistics for those school districts, spanning from 2013 to 2017, are shown in Table 35. Of these 17 high-science-achievement school districts, 76 percent offered comprehensive science courses, and 24 percent offered subject-specific science courses, identical to the percentages for all school districts. The high science-achievement school districts had smaller mean student populations, smaller mean percentages of low SES students, and smaller mean percentages of ELs than all school districts collectively. The high science-achievement school districts had higher mean scale scores in science for all three groups of students than all school districts collectively.

Table 35

Highest Science-achievement School Districts[*]*, 2013–2017* (*N* = 17 school districts)

	School district group		Student population			Mean scale score			Pass
				Low SES	EL	All	Low		rate
Statistic	C	S	Total	%	%	students	SES	EL	%
M	13	4	25,139	49.53	2.94	205.31	195.90	182.91	57.62
Mdn	-	-	15,169	47.79	2.29	204.77	198.67	182.05	57.65
SD	-	-	23,895	9.92	2.42	1.88	5.30	4.92	3.97
Min	-	-	1,429	23.19	0.63	203.49	186.51	173.82	52.40
Max	-	-	72,159	64.61	8.69	209.38	202.38	191.85	66.20

Note. [*]School districts with 2013–2017 mean scale scores for all students of 203 or higher.

Of all 67 school districts, 19 had 2013–2017 mean scale scores below 197 (the cut-point for the first quartile). Of these 19 low-science-achievement school districts, 95 percent offered comprehensive, and five percent offered subject-specific science courses. The low science-achievement school districts had greater mean student populations, low SES student percentages, and EL percentages than the overall school district means. The low-science-achievement school districts included 4 of the state's 10 largest school districts (including the two largest). The low-science-achievement school districts also included 9 of the state's 10 highest poverty (by low SES percentage) school districts, and 4 of the state's 10 highest EL percentage school districts.

Table 36

*Lowest Science-achievement School Districts**, *2013–2017* (*N* = 19 school districts)

	School district group		Student population			Mean scale score			Pass rate
Statistic	C	S	Total	Low SES %	EL %	All students	Low SES	EL	%
M	18	1	52,107	66.47	7.74	190.91	186.35	175.75	32.00
Mdn	-	-	6,479	69.68	7.62	193.63	185.06	177.17	34.26
SD	-	-	98,987	8.70	5.34	9.40	3.67	5.34	7.94
Min	-	-	889	48.76	0.64	155.80	182.43	161.73	17.40
Max	-	-	356,659	80.77	20.72	196.83	194.12	182.27	40.91

Note. * School districts with 2013–2017 mean scale scores for all students below 197.

Figure 6 shows the 8th grade FCAT 2.0 Science/SSA school district mean scale scores for 2013 to 2017, for the three student groups in the two school district groups. Achievement gaps for low SES students and ELs are evident in the figure, and well-supported by the literature (Becker & Luthar, 2002; Cosentino de Cohen et al., 2005; Hanushek, 2010; Hattie, 2009; Ladd, 2012; Miller et al., 2013; NCES, 2016; Wiseman, 2012).

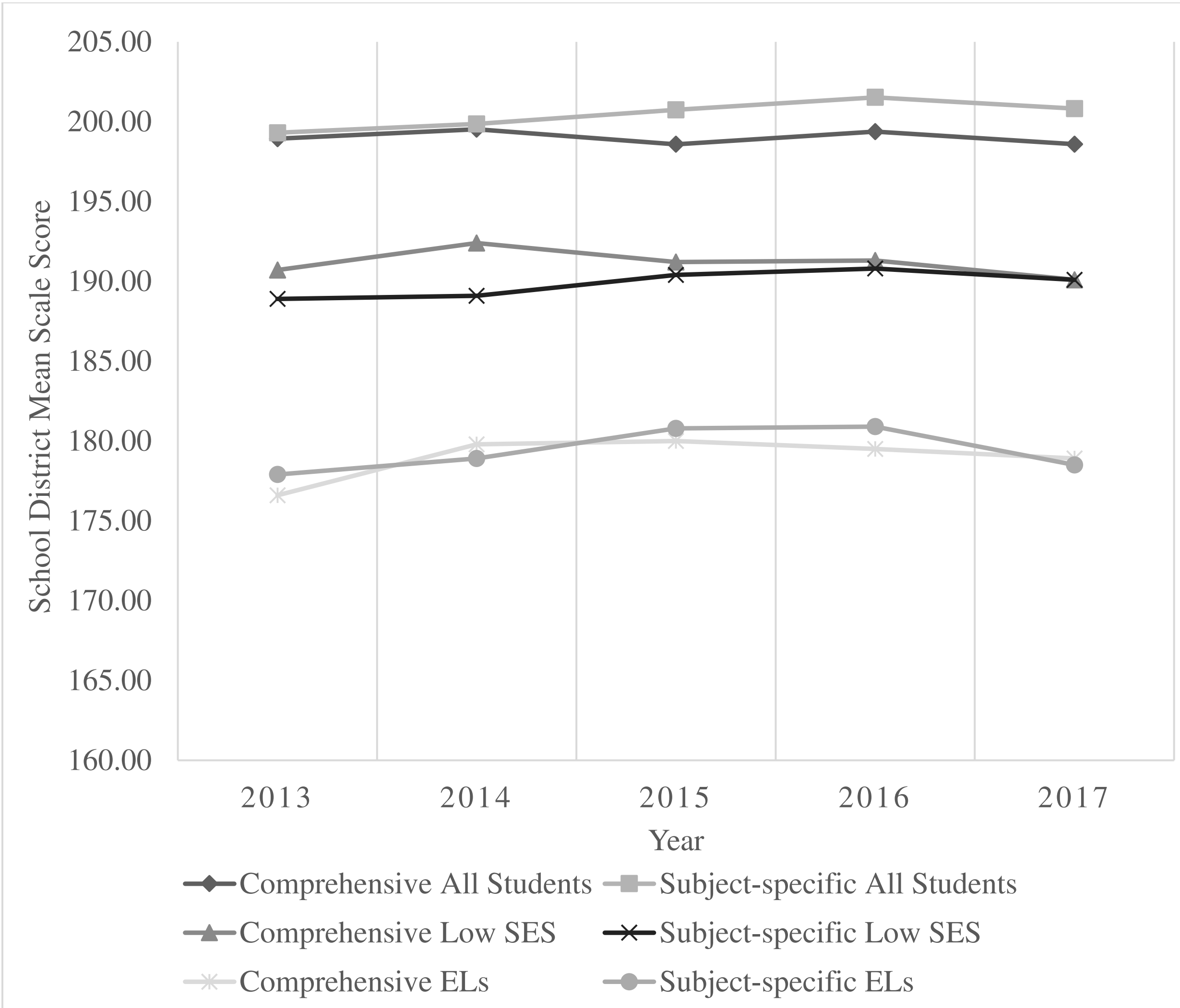

Figure 6. Independent samples, 8th Grade FCAT 2.0 Science/SSA school district mean scale scores 2013–2017 for all students, low SES students, and ELs in comprehensive and subject-specific school district groups.

Independent samples *t*-tests were conducted to quantify differences in the 2013–2017 8th grade FCAT 2.0 Science/SSA school district mean scale scores between student groups (all students vs. low SES students, and all students vs. ELs). As shown in Table 37, the effect size of the difference in school district mean scale scores between all students and low SES students is very large (Sawilowsky, 2003). The effect size of the difference in school district mean scale scores between all students and ELs is huge (Sawilowsky, 2003). While achievement gaps were not the focus of this study, they were the reason for the separate analyses of school district mean scale scores for these three student groups. The focus of this study was to determine if there were significant differences in FCAT 2.0 Science/SSA achievement of the three student groups, between the comprehensive and subject-specific school district groups. The test results for the four research questions are presented next, followed by the results of six additional analyses.

Table 37

Differences in School District Mean Scale Scores Between Student Groups

	Descriptive statistics						Paired samples tests					
										95% CI		
Student group	*N*	*M*	*SD*	*SEM*	Skew	Kurtosis	*t*	*df*	*p*	*LL*	*UL*	*d*
All students	334	199.34	5.59	0.31	-0.86	1.57	28.05	333	<.001	7.81	8.98	1.54
Low SES	334	190.94	6.26	0.34	0.28	-0.86						
All students	179	199.78	4.15	0.31	-0.86	1.57	50.40	178	<.001	19.95	21.57	3.77
ELs	179	179.02	5.30	0.40	0.19	0.36						

Research Question 1

Research Question 1: To what extent did 8th grade FCAT 2.0 Science/SSA school district mean scale scores for 2013, 2014, 2015, 2016, and 2017 differ between two groups of Florida school districts: those that offered comprehensive science courses and those that offered subject-specific science courses?

To answer this question, independent samples *t*-tests were used to test differences in each year's 8th grade FCAT 2.0 Science/SSA school district mean scale scores for three groups of students: all students, low SES students, and ELs; between the two groups of school districts.

No controls were used for school district population, percentage of low SES students, percentage of ELs, or any other demographic variable. Sample sizes of the school district groups (*n*) were unequal because, from 2013 to 2017, most Florida school districts offered comprehensive science courses. Skewness and kurtosis were examined to test the assumption of normality of distribution. Levene's test was used to test the assumption of equality of variance of the samples. Effect size (Cohen's *d*) was calculated for statistically significant results.

Independent Samples Tests of School District Mean Scale Scores for All Students

2013 (All Students)

As shown in Table 38, the 2013 8th grade FCAT 2.0 Science/SSA school district mean scale score for all students was numerically lower for the comprehensive group than for the subject-specific group. The mean scale score distributions for both groups were sufficiently normal for the purposes of conducting a *t*-test, i.e., skew < |2.0| and kurtosis < |9.0| (Schmider et al., 2010). The assumption of equality of variances was satisfied by Levene's test, $F(65) = 0.44$, $p = .51$ (George & Mallery, 2010). The independent samples *t*-test revealed no statistically significant difference in 2013 8th grade FCAT 2.0 Science/SSA school district mean scale score for all students between the comprehensive (C) and subject-specific (S) groups of school districts.

Table 38

2013 Independent Samples Tests of School District Mean Scale Scores for All Students (N = 67 school districts)

School district group	Descriptive statistics							Independent samples test				
											95% CI	
	n	*M*	*Mdn*	*SD*	*SEM*	Skew	Kurtosis	*t*	*df*	*p*	*LL*	*UL*
C	49	198.94	199.00	5.84	0.83	-0.76	0.97	-0.23	65	.82	-3.60	2.87
S	18	199.31	199.56	5.98	1.41	-1.47	4.65					

2014 (All Students)

As shown in Table 39, the 2014 FCAT 2.0 Science/SSA mean scale score for all students was numerically lower for the comprehensive group than for the subject-specific group. The distributions of school district mean scale scores for all students in both groups were sufficiently normal for the purposes of conducting a *t*-test, i.e., skew < |2.0| and kurtosis < |9.0| (Schmider et al., 2010). Levene's test confirmed the assumption of equality of variances, $F(65) = 0.61$, $p = .44$ (George & Mallery, 2010). The independent samples *t*-test, shown in Table 39, revealed no statistically significant difference in the 2014 8th grade FCAT 2.0 Science/SSA school district mean scale score for all students between the comprehensive and subject-specific school district groups.

Table 39

2014 Independent Samples Tests of School District Mean Scale Scores for All Students (N = 67 school districts)

School district group	Descriptive statistics							Independent samples test				
											95% CI	
	n	*M*	*Mdn*	*SD*	*SEM*	Skew	Kurtosis	*t*	*df*	*p*	*LL*	*UL*
C	52	199.52	200.11	5.23	0.73	-0.68	1.06	-0.23	65	.82	-3.29	2.61
S	15	199.86	199.25	4.26	1.10	-0.13	-0.69					

2015 (All Students)

As shown in Table 40, the 2015 FCAT 2.0 Science/SSA mean scale score for all students was numerically lower for the comprehensive group than for the subject-specific group. The mean scale score distributions for both school district groups were sufficiently normal for the purposes of conducting a *t*-test, i.e., skew < |2.0| and kurtosis < |9.0| (Schmider et al., 2010). Levene's test did not confirm the assumption of equality of variances, $F(65) = 5.25$, $p = .03$ (George & Mallery, 2010), so the degrees of freedom were reduced from $df = 65$ to $df = 52.73$. While the independent samples *t*-test results (shown in Table 40) showed no statistically significant difference in 2015 8th grade FCAT 2.0 Science/SSA school district mean scale score for all students between the two school district groups, the difference did approach statistical significance, as the *p*-value was just over the $\alpha = .05$ threshold.

Table 40

2015 Independent Samples Tests of School District Mean Scale Scores for All Students (*N* = 67 school districts)

School district group	Descriptive statistics							Independent samples test				
											95% CI	
	n	*M*	*Mdn*	*SD*	*SEM*	Skew	Kurtosis	*t*	*df*	*p*	*LL*	*UL*
C	53	198.58	199.90	6.11	0.84	-1.01	2.23	-2.00	52.73	.051	-4.32	0.01
S	14	200.73	201.20	2.53	0.68	-0.13	-0.50					

2016 (All Students)

Table 41 shows that the 2016 FCAT 2.0 Science/SSA school district mean scale score for all students was numerically lower for the comprehensive group than for the subject-specific group. The 2016 FCAT 2.0 Science/SSA mean scale score distributions for both groups met the assumption of normality for the purposes of conducting a *t*-test, i.e., skew < |2.0| and kurtosis < |9.0|; (Schmider et al., 2010). The assumption of equality of variances was not satisfied by Levene's test, $F(65) = 4.26$, $p = .04$ (George & Mallery, 2010), so the degrees of freedom were reduced from $df = 65$ to $df = 46.02$. Also shown in Table 41, the independent samples *t*-test revealed no statistically significant difference in the 2016 8th grade FCAT 2.0 Science/SSA school district mean scale score for all students between the comprehensive and subject-specific school district groups. The difference did approach statistical significance, as the *p*-value was just over the $\alpha = .05$ threshold.

Table 41

2016 Independent Samples Tests of School District Mean Scale Scores for All Students (*N* = 67 school districts)

School district group	Descriptive statistics							Independent samples test				
											95% CI	
	n	*M*	*Mdn*	*SD*	*SEM*	Skew	Kurtosis	*t*	*df*	*p*	*LL*	*UL*
C	52	199.38	199.43	5.83	0.81	-0.35	0.22	-1.91	46.02	.062	-4.40	0.11
S	15	201.52	200.92	3.01	0.78	0.18	-0.73					

2017 (All Students)

As shown in Table 42, the 2017 FCAT 2.0 Science/SSA mean scale score for all students was numerically lower for the comprehensive group than for the subject-specific group. The mean scale score distributions for both groups met the assumption of normality, i.e., skew < |2.0| and kurtosis < |9.0| (Schmider et al., 2010). The assumption of equality of variances was not satisfied by Levene's test, $F(65) = 5.82$, $p = .02$ (George & Mallery, 2010), so the degrees of freedom were reduced from $df = 65$ to $df = 57.49$. Also shown in Table 42, the independent samples *t*-test revealed no statistically significant difference in the 2017 FCAT 2.0 Science/SSA school district mean scale score for all students between the comprehensive and subject-specific school district groups. The difference did approach statistical significance, as the *p*-value was just over the $\alpha = .05$ threshold.

Table 42

2017 Independent Samples Tests of School District Mean Scale Scores for All Students (N = 66 school districts*)

School district group	Descriptive statistics							Independent samples test				
											95% CI	
	n	M	Mdn	SD	SEM	Skew	Kurtosis	t	df	p	LL	UL
C	49	198.59	199.00	6.85	0.98	-0.71	0.54	-1.77	57.49	.083	-4.74	0.30
S	17	200.81	200.63	3.26	0.79	0.40	0.65					

Note. *One school district reported no 8th grade FCAT 2.0 Science/SSA results for 2017

Independent Samples Tests of School District Mean Scale Scores for Low SES Students

2013 (Low SES Students)

The 2013 8th grade FCAT 2.0 Science/SSA school district mean scale score for low SES students was numerically higher for the comprehensive group than for the subject-specific group, as shown in Table 43. The distributions of both school district groups of school districts were sufficiently normal for the purposes of conducting a t-test, i.e., skew < |2.0| and kurtosis < |9.0| (Schmider et al., 2010). The assumption of equality of variances was not satisfied by Levene's test, $F(65) = 8.44$, $p = .01$ (George & Mallery, 2010), so degrees of freedom were reduced from $df = 65$ to $df = 48.30$. Also shown in Table 43, the independent samples t-test revealed no statistically significant difference in the 2013 8th grade FCAT 2.0 Science/SSA school district mean scale score for low SES students between the comprehensive and subject-specific groups.

Table 43

2013 Independent Samples Tests of School District Mean Scale Scores for Low SES Students (*N* = 67 school districts)

School district group	Descriptive statistics							Independent samples test				
											95% CI	
	n	*M*	*Mdn*	*SD*	*SEM*	Skew	Kurtosis	*t*	*df*	*p*	*LL*	*UL*
C	49	190.67	190.07	7.42	1.06	0.14	-1.09	1.18	48.30	.25	-1.28	4.88
S	18	188.87	189.05	4.68	1.10	0.32	-0.23					

2014 (Low SES Students)

As shown in Table 44, the 2014 8th grade FCAT 2.0 Science/SSA mean scale score for low SES students was numerically higher for the comprehensive group than for the subject-specific group. The distributions of 2014 FCAT 2.0 Science/SSA mean scale scores for low SES students in both school district groups were sufficiently normal for the purposes of conducting a *t*-test, i.e., skew < |2.0| and kurtosis < |9.0| (Schmider et al., 2010). The assumption of equality of variances was satisfied by Levene's test, $F(65) = 2.37$, $p = .13$ (George & Mallery, 2010). The independent samples *t*-test, shown in Table 44, revealed no statistically significant difference in the 2014 8th grade FCAT 2.0 Science/SSA school district mean scale score for low SES students between the two school district groups. The *p*-value was just over the $\alpha = .05$ threshold, indicating that the difference approached statistical significance.

Table 44

2014 Independent Samples Tests of School District Mean Scale Scores for Low SES Students (*N* = 67 school districts)

School district group	Descriptive statistics							Independent samples test				
											95% CI	
	n	*M*	*Mdn*	*SD*	*SEM*	Skew	Kurtosis	*t*	*df*	*p*	*LL*	*UL*
C	52	192.39	192.03	6.04	0.84	0.13	-1.06	1.95	65	.055	-0.08	6.70
S	15	189.08	186.95	4.75	1.23	1.27	1.56					

2015 (Low SES Students)

As shown in Table 45, the 2015 8th grade FCAT 2.0 Science/SSA mean scale score for low SES students was numerically higher for the comprehensive group than for the subject-specific group. The distributions of school district mean scale scores for both groups were sufficiently normal for the purposes of conducting a *t*-test, i.e., skew < |2.0| and kurtosis < |9.0| (Schmider et al., 2010). The assumption of equality of variances was satisfied by Levene's test, $F(65) = 2.85$, $p = .10$ (George & Mallery, 2010). The independent samples *t*-test, also shown in Table 45, revealed no statistically significant difference in the 2015 8th grade FCAT 2.0 Science/SSA school district mean scale score for low SES students between the comprehensive and subject-specific groups of school districts.

Table 45

2015 Independent Samples Tests of School District Mean Scale Scores for Low SES Students (*N* = 67 school districts)

School district group	Descriptive statistics							Independent samples test				
											95% CI	
	n	*M*	*Mdn*	*SD*	*SEM*	Skew	Kurtosis	*t*	*df*	*p*	*LL*	*UL*
C	53	191.20	190.79	6.62	0.91	-0.05	-0.61	0.41	65	.68	-2.99	4.54
S	14	190.43	188.57	4.66	1.24	0.56	-1.41					

2016 (Low SES Students)

As shown in Table 46, the 2016 8th grade FCAT 2.0 Science/SSA mean scale score for low SES students was numerically higher for the comprehensive group than for the subject-specific group. The distributions of 2016 FCAT 2.0 Science/SSA mean scale scores for low SES students in both groups were sufficiently normal for the purposes of conducting a *t*-test, i.e., skew < |2.0| and kurtosis < |9.0| (Schmider et al., 2010). The assumption of equality of variances was satisfied by Levene's test, $F(65) = 2.63$, $p = .11$ (George & Mallery, 2010). The independent samples *t*-test, also shown in Table 46, revealed that there was not a statistically significant difference in the 2016 8th grade FCAT 2.0 Science/SSA mean scale score for low SES students between the comprehensive and subject-specific school district groups.

Table 46

2016 Independent Samples Tests of School District Mean Scale Scores for Low SES Students (*N* = 67 school districts)

School district group	Descriptive statistics							Independent samples test				
											95% CI	
	n	*M*	*Mdn*	*SD*	*SEM*	Skew	Kurtosis	*t*	*df*	*p*	*LL*	*UL*
C	52	191.32	190.18	6.54	0.91	0.37	-1.11	0.26	65	.80	-3.19	4.13
S	15	190.85	188.87	5.05	1.30	0.93	-0.05					

2017 (Low SES Students)

Table 47 shows that the 2017 8th grade FCAT 2.0 Science/SSA school district mean scale score for low SES students was numerically higher for the comprehensive group than for the subject-specific group. The distributions of mean scale scores for low SES students for both groups were sufficiently normal for the purposes of conducting a *t*-test, i.e., skew < |2.0| and kurtosis < |9.0| (Schmider et al., 2010). The assumption of equality of variances was satisfied by Levene's test, *F*(64) = 1.02, *p* = .32 (George & Mallery, 2010). The independent samples *t*-test, also shown in Table 47, revealed no statistically significant difference in 2017 8th grade FCAT 2.0 Science/SSA school district mean scale score for low SES students between the comprehensive and subject-specific groups of school districts.

Table 47

2017 Independent Samples Tests of School District Mean Scale Scores for Low SES Students (*N* = 66 school districts*)

School district group	Descriptive statistics							Independent samples test				
											95% CI	
	n	*M*	*Mdn*	*SD*	*SEM*	Skew	Kurtosis	*t*	*df*	*p*	*LL*	*UL*
C	49	190.78	189.06	6.44	0.92	0.41	-1.04	0.40	64	.69	-2.81	4.21
S	17	190.08	187.96	5.58	1.35	0.74	-1.10					

Note. *One school district reported no 8th grade FCAT 2.0 Science/SSA results for 2017.

Independent Samples Tests of School District Mean Scale Scores for ELs

2013 (ELs)

As shown in Table 48, the 2013 8th grade FCAT 2.0 Science/SSA mean scale score for ELs was numerically lower for the comprehensive group than for the subject-specific group. The distributions of both groups of school districts were sufficiently normal for the purposes of conducting a *t*-test, i.e., skew < |2.0| and kurtosis < |9.0| (Schmider et al., 2010). Levene's test, $F(31) = 1.03$, $p = .32$, confirmed the assumption of equality of variances (George & Mallery, 2010). Also shown in Table 48, the independent samples *t*-test revealed no statistically significant difference in the 2013 8th grade FCAT 2.0 Science/SSA school district mean scale score for ELs between two groups of school districts.

Table 48

2013 Independent Samples Tests of School District Mean Scale Scores for ELs (*N* = 33 school districts*)

School district group	Descriptive statistics							Independent samples test			95% CI	
	n	*M*	*Mdn*	*SD*	*SEM*	Skew	Kurtosis	*t*	*df*	*p*	*LL*	*UL*
C	22	175.66	176.80	5.16	1.10	-0.88	1.37	-1.27	31	.21	-5.83	1.36
S	11	177.90	178.78	3.85	1.16	-1.51	2.92					

Note. *N is less than the total number of Florida school districts due to data suppression for school districts with fewer than 10 ELs assessed (Florida Department of Education [FLDOE], 2017i).

2014 (ELs)

As shown in Table 49, the 2014 8th grade FCAT 2.0 Science/SSA mean scale score for ELs was numerically higher for the comprehensive group than for the subject-specific group. The distributions of 2014 FCAT 2.0 Science/SSA mean scale scores for ELs in school districts that offered either type of science course were sufficiently normal for the purposes of conducting a *t*-test, i.e., skew < |2.0| and kurtosis < |9.0| (Schmider et al., 2010). Levene's test, $F(31) = 0.13$, $p = .72$, confirmed the assumption of equality of variances (George & Mallery, 2010). Also shown in Table 49, the independent samples *t*-test revealed no statistically significant difference in 2014 8th grade FCAT 2.0 Science/SSA school district mean scale scores for ELs between the two school district groups.

Table 49

2014 Independent Samples Tests of School District Mean Scale Scores for ELs (N = 33 school districts*)

School district group	Descriptive statistics							Independent samples test				
											95% CI	
	n	*M*	*Mdn*	*SD*	*SEM*	Skew	Kurtosis	*t*	*df*	*p*	*LL*	*UL*
C	22	179.82	179.88	5.63	1.20	0.80	2.18	0.47	31	.64	-3.14	5.00
S	11	178.89	177.06	4.90	1.48	1.67	3.85					

Note. *N is less than the total number of Florida school districts due to data suppression for school districts with fewer than 10 ELs assessed (FLDOE, 2017e)

2015 (ELs)

As shown in Table 50, the 2015 8th grade FCAT 2.0 Science/SSA mean scale score for ELs was numerically lower for the comprehensive group than for the subject-specific group. The distributions of 2015 FCAT 2.0 Science/SSA mean scale scores for ELs for both groups were sufficiently normal for the purposes of conducting a t-test, i.e., skew < |2.0| and kurtosis < |9.0| (Schmider et al., 2010). Levene's test, $F(35) = 8.40$, $p = .01$, did not confirm the assumption of equality of variances (George & Mallery, 2010), so the degrees of freedom for the independent samples t-test were reduced from $df = 35$ to $df = 34.84$. Also shown in Table 50, the independent samples t-test revealed no statistically significant difference in the 2015 8th grade FCAT 2.0 Science/SSA school district mean scale score for ELs between the comprehensive and subject-specific school district groups.

Table 50

2015 Independent Samples Tests of School District Mean Scale Scores for ELs (*N* = 37 school districts*)

School district group	Descriptive statistics							Independent samples test				
											95% CI	
	n	*M*	*Mdn*	*SD*	*SEM*	Skew	Kurtosis	*t*	*df*	*p*	*LL*	*UL*
C	27	180.05	179.94	6.03	1.16	-0.26	-0.37	-0.52	34.84	.61	-3.50	2.08
S	10	180.76	180.89	2.33	0.74	-0.40	1.35					

Note. *N is less than the total number of Florida school districts due to data suppression for school districts with fewer than 10 ELs assessed (FLDOE, 2017e)

2016 (ELs)

As shown in Table 51, the 2016 8th grade FCAT 2.0 Science/SSA mean scale score for ELs was numerically lower for the comprehensive group than for the subject-specific group. The distributions of mean scale scores for ELs in both groups were sufficiently normal for the purposes of conducting a *t*-test, i.e., skew < |2.0| and kurtosis < |9.0| (Schmider et al., 2010). Levene's test, $F(38) = 0.35$, $p = .56$, confirmed the assumption of equality of variances (George & Mallery, 2010). Also shown in Table 51, the independent samples *t*-test showed no statistically significant difference in the 2016 8th grade FCAT 2.0 Science/SSA school district mean scale score for ELs between the comprehensive and subject-specific groups of school districts.

Table 51

2016 Independent Samples Tests of School District Mean Scale Scores for ELs (*N* = 40 school districts*)

School district group	Descriptive statistics							Independent samples test				
											95% CI	
	n	*M*	*Mdn*	*SD*	*SEM*	Skew	Kurtosis	*t*	*df*	*p*	*LL*	*UL*
C	29	179.53	178.00	5.90	1.10	0.66	-0.50	-0.66	38	.51	-5.43	2.77
S	11	180.86	181.53	5.20	1.57	-0.46	-0.10					

Note. **N* is less than the total number of Florida school districts due to data suppression for school districts with fewer than 10 ELs assessed (FLDOE, 2017e)

2017 (ELs)

As shown in Table 52, the 2017 8th grade FCAT 2.0 Science/SSA school district mean scale score for ELs was numerically higher for the comprehensive group than for the subject-specific group. The distributions of mean scale scores for ELs for both groups were sufficiently normal for the purposes of conducting a *t*-test, i.e., skew < |2.0| and kurtosis < |9.0| (Schmider et al., 2010). Levene's test, $F(34) = 0.56$, $p = .46$, confirmed the assumption of equality of variances (George & Mallery, 2010). Also shown in Table 52, the independent samples *t*-test showed no statistically significant difference in the 2017 8th grade FCAT 2.0 Science/SSA school district mean scale score for ELs between the comprehensive and subject-specific school district groups.

Table 52

2017 Independent Samples Tests of School District Mean Scale Scores for ELs (N = 36 school districts*)

School district group	Descriptive statistics							Independent samples test				
											95% CI	
	n	*M*	*Mdn*	*SD*	*SEM*	Skew	Kurtosis	*t*	*df*	*p*	*LL*	*UL*
C	25	178.85	178.53	5.06	1.01	0.84	0.35	0.19	34	.85	-3.16	3.83
S	11	178.52	180.08	3.91	1.18	-1.03	1.14					

Note. *N is less than the total number of Florida school districts due to data suppression for school districts with fewer than 10 ELs assessed (FLDOE, 2017e).

Summary of Research Question 1 Independent Samples Tests

Research Question 1 used independent samples *t*-tests and all available school district 8th grade FCAT 2.0 Science/SSA school district mean scale scores to test for differences in the school district mean scale scores for three groups of students (all students, low SES students, and ELs), between two groups of school districts (those that offered comprehensive and those that offered subject-specific science courses). The distributions of each year's school district mean scale scores for all students in both groups of school districts satisfied the tests for normality of distribution. The 2015, 2016, and 2017 school district mean scale scores for all students did not satisfy the test for equality of variance, so the degrees of freedom were adjusted accordingly in the independent samples *t*-tests.

Mean Scale Scores for All Students

Although the school district mean scales scores for all students were numerically greater each year for the subject-specific group than for the comprehensive, no statistically significant differences were found in the 2013, 2014, 2015, 2016, or 2017 school district mean scale scores for all students, between the two groups of school districts. The tests of the 2015, 2016, and 2017 school district mean scale scores for all students (which had reduced degrees of freedom due to unequal variance) approached statistical significance, with *p*-values just over the $\alpha = .05$ threshold.

As shown in Table 53, the dispersion of school district mean scale scores for all students for the comprehensive group was greater than for the subject-specific group in 2014 through 2017. This would have held true for 2013 as well, if not for a single low outlier in the 2013 subject-specific group. The boxplot in Figure 7 more clearly depicts these differences in dispersion, as well as the presence of the outliers that impacted the Levene's test for equal variance. All but one of the low outliers were in the comprehensive group. Fictitious school district names shown in the boxplot show that five school districts comprised the low outliers, three of which were low outliers in multiple years. One school district, Tarek, was a low outlier before and after changing from subject-specific to comprehensive science courses in 2014. Student populations in all four low outlier school districts ranged from 889 to 99,656 students. Of these students, 62 to 72 percent were low SES students, and 1 to 7 percent were ELs.

Four school districts comprised the five high outliers. All were in the subject-specific group. None exceeded the maximum mean scale score of the comprehensive group. Student populations of the high outlier school districts ranged from 8,468 to 66,132 students. Low SES percentages ranged from 43 to 49 percent, and EL percentages ranged from 2 to 9 percent. One school district, Renato, appeared twice as a high outlier. The Renato school district student population (30,350 students), percentage of low SES students (43%), and percentage of ELs (2.95%) were all less than the means of these demographics for all 67 school districts.

The boxplot shows that the median school district mean scale score for all students in the subject-specific group was higher than that of the comprehensive group for four of the five years. The median school district mean scale score for all students in neither group reached the minimum passing score in any of the five years. The third quartile for both school district groups hovered at or below the minimum passing score each year.

Table 53

2013–2017 Independent Samples Tests of School District Mean Scale Scores for All Students

		Descriptive statistics					Independent samples *t*-tests				
										95% CI	
Year	School district group	*n*	*M*	*Mdn*	*SD*	*SEM*	*t*	*df*	*p*	*LL*	*UL*
2013	C	49	198.94	199.00	5.84	0.83	-0.23	65	.82	-3.60	2.87
	S	18	199.31	199.56	5.98	1.41					
2014	C	52	199.52	200.11	5.23	0.73	-0.23	65	.82	-3.29	2.61
	S	15	199.86	199.25	4.26	1.10					
2015	C	53	198.58	199.90	6.11	0.84	-2.00	52.73	.051	-4.32	0.01
	S	14	200.73	201.20	2.53	0.68					
2016	C	52	199.38	199.43	5.83	0.81	-1.91	46.02	.062	-4.40	0.11
	S	15	201.52	200.92	3.01	0.78					
2017	C	49	198.59	199.00	6.85	0.98	-1.77	57.49	.083	-4.74	0.30
	S	17	200.81	200.63	3.26	0.79					

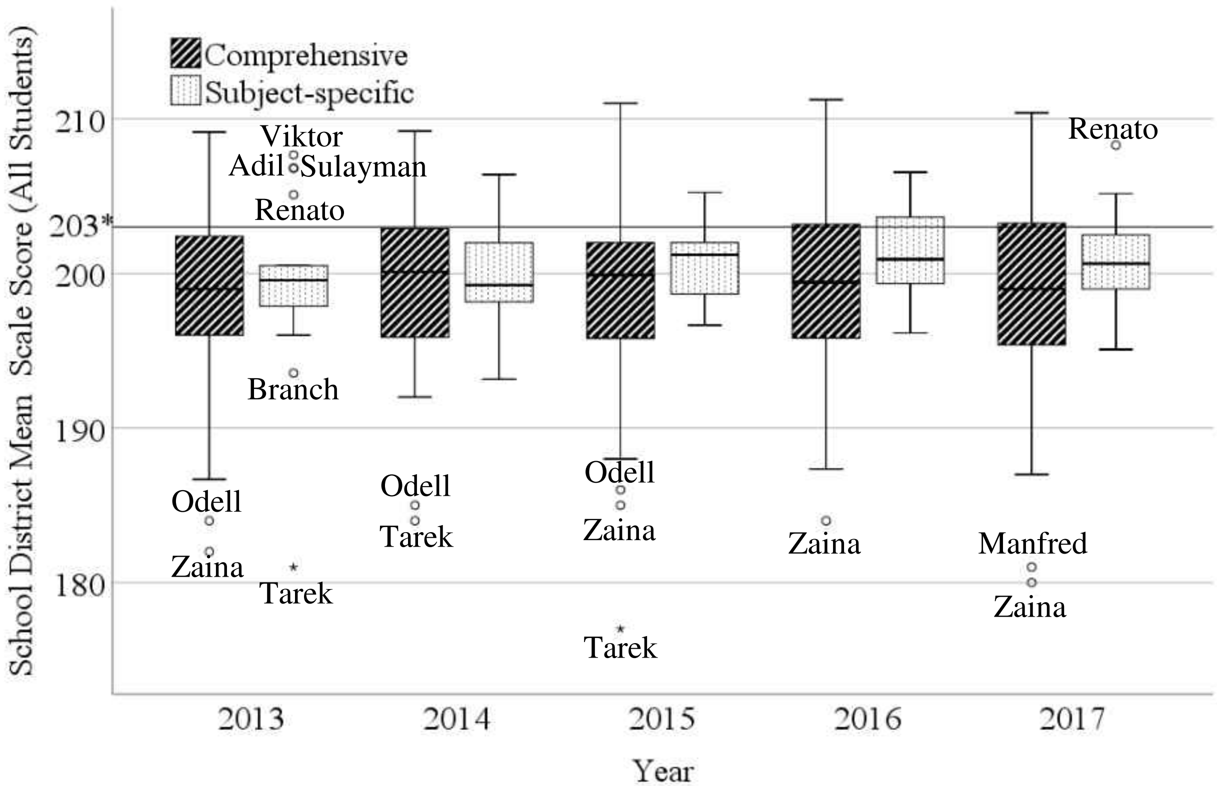

Figure 7. Boxplot of 2013–2017 school district mean scale scores for all students, by school district group. Independent samples. *Minimum passing score (FLDOE, 2017o). Circles represent outliers; asterisks represent extreme outliers with values more than three times the height of the boxes (George & Mallery, 2010). Fictitious school district names are shown for the outliers.

Mean Scale Scores for Low SES Students

Although the school district mean scale scores for low SES students were numerically higher each year for the comprehensive group than for the subject-specific group, no statistically significant differences were found in the 2013, 2014, 2015, 2016, or 2017 school district mean scale scores for low SES students between the two groups of school districts. The tests of the 2014 school district mean scale scores for low SES students approached statistical significance, with a *p*-value just over the $\alpha = .05$ threshold. Greater dispersion in the comprehensive group mean scale scores was evident, as shown in Table 54 and the boxplot in Figure 8. No low outliers were evident. Only one high outlier appeared, Renato school district, in the subject-specific group. This school district was also a subject-specific group high outlier for two years in the school district mean scale scores for all students.

The boxplot reveals that the median school district mean scale score for low SES students was higher each year for the comprehensive group than for the subject-specific group. The maximum school district mean scale scores for low SES students in the subject-specific group did not reach the minimum passing score of 203 in any year from 2013 to 2017. For the comprehensive group, the maximum school district mean scale score for low SES students was just above the minimum passing score in 2013, 2014, 2016, and 2017, and just below the minimum passing score in 2015.

Table 54

2013–2017 Independent Samples Tests of School District Mean Scale Scores for Low SES Students

		Descriptive statistics					Independent samples *t*-tests				
										95% CI	
Year	School district group	*n*	*M*	*Mdn*	*SD*	*SEM*	*t*	*df*	*p*	*LL*	*UL*
2013	C	49	190.67	190.07	7.42	1.06	1.18	48.30	.25	-1.28	4.88
	S	18	188.87	189.05	4.68	1.10					
2014	C	52	192.39	192.03	6.04	0.84	1.95	65	.055	-0.08	6.70
	S	15	189.08	186.95	4.75	1.23					
2015	C	53	191.20	190.79	6.62	0.91	0.41	65	.68	-2.99	4.54
	S	14	190.43	188.57	4.66	1.24					
2016	C	52	191.32	190.18	6.54	0.91	0.26	65	.80	-3.19	4.13
	S	15	190.85	188.87	5.05	1.30					
2017	C	49	190.78	189.06	6.44	0.92	0.40	64	.69	-2.81	4.21
	S	17	190.08	187.96	5.58	1.35					

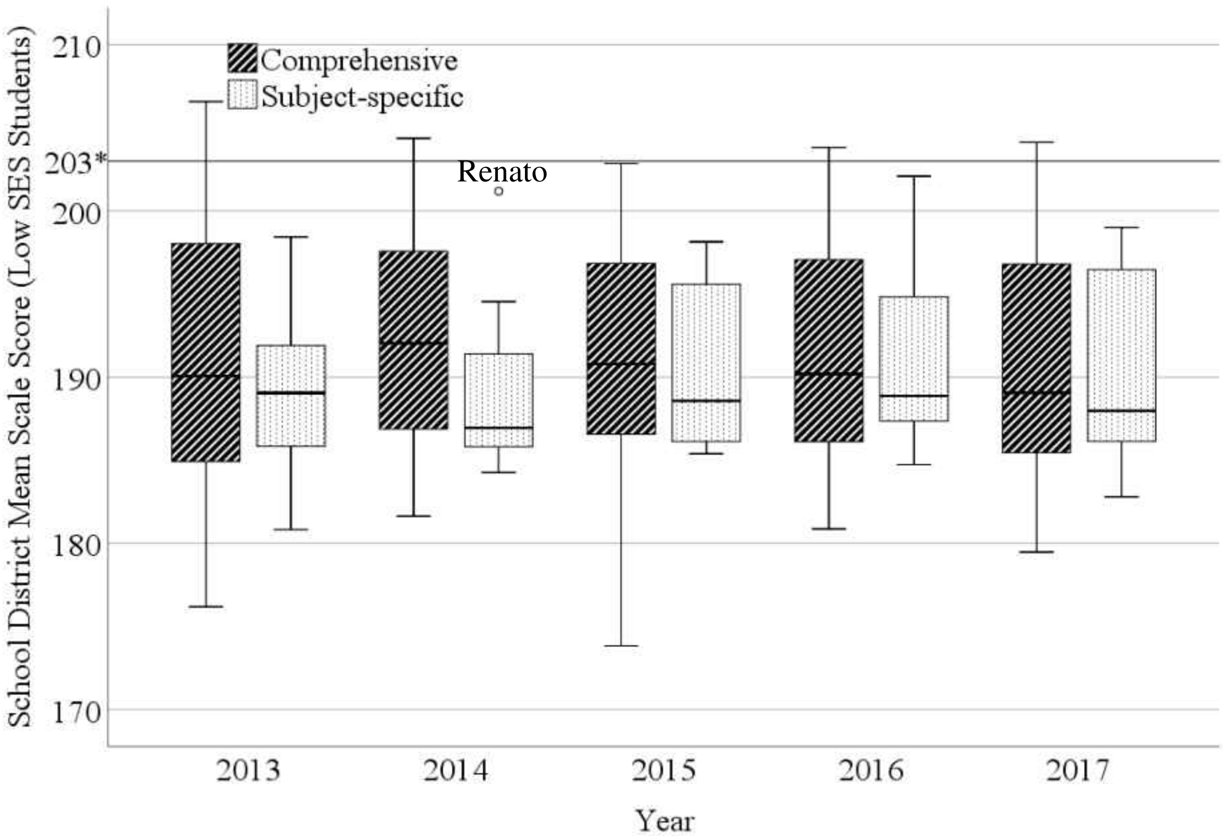

Figure 8. Boxplot of 2013–2017 school district mean scale scores for low SES students, by school district group. Independent samples. *Minimum passing score (FLDOE, 2017o).

Mean Scale Scores for ELs

Although the subject-specific group mean scale scores for ELs were numerically higher in three of the five years, no statistically significant differences were found in the 2013, 2014, 2015, 2016, or 2017 school district mean scale scores for ELs between the two groups of school districts. None of the tests approached statistical significance. Greater dispersion and several outliers were evident in the comprehensive group school district mean scale scores for ELs, as shown in Table 55 and Figure 9.

Four school districts comprised the five low outliers. Three were in the subject-specific group. The low outlier school districts ranged from 4,851 to 66,133 in student population. Low SES student percentage ranged from 46.55 to 69.68 percent. Percentage of ELs ranged from 4.30 to 14.34 percent. Only one low outlier, Viktor school district, appeared twice. In 2013, Viktor school district was both a low outlier in the school district mean scale scores for ELs, and a high outlier in the mean scale scores for all students. The student population (8,468 students), low SES percentage (47.17%), and EL percentage (8.69) of Viktor school district were less than the overall means for all 67 school districts.

The four high outliers were comprised of four school districts, ranging in student population from 28,648 to 36,103 students, percentage of low SES students from 40.72 to 49.21 percent, and percentage of ELs from 1.79 to 2.95 percent. Renato school district again appeared as a high outlier in the subject-specific group. The median school district mean scale score for ELs in the subject-specific group was higher than that of the comprehensive group in four of the five years. The maximum school district mean scale score for ELs in neither group reached the minimum passing score.

Table 55

2013–2017 Independent Samples Tests of School District Mean Scale Scores for ELs

							Independent samples *t*-tests				
	School district	Descriptive statistics								95% CI	
Year	group	*n*	*M*	*Mdn*	*SD*	*SEM*	*t*	*df*	*p*	*LL*	*UL*
2013	C	22	175.66	178.78	5.16	1.10	-1.27	31	.21	-5.83	1.36
	S	11	177.90	176.80	3.85	1.16					
2014	C	22	179.82	179.88	5.63	1.20	0.47	31	.64	-3.14	5.00
	S	11	178.89	177.06	4.90	1.48					
2015	C	27	180.05	179.94	6.03	1.16	-0.52	34.84	.61	-3.50	2.08
	S	10	180.76	180.89	2.33	0.74					
2016	C	29	179.53	178.00	5.90	1.10	-0.66	38	.51	-5.43	2.77
	S	11	180.86	181.53	5.20	1.57					
2017	C	25	178.85	178.53	5.06	1.01	0.19	34	.85	-3.16	3.83
	S	11	178.52	180.08	3.91	1.18					

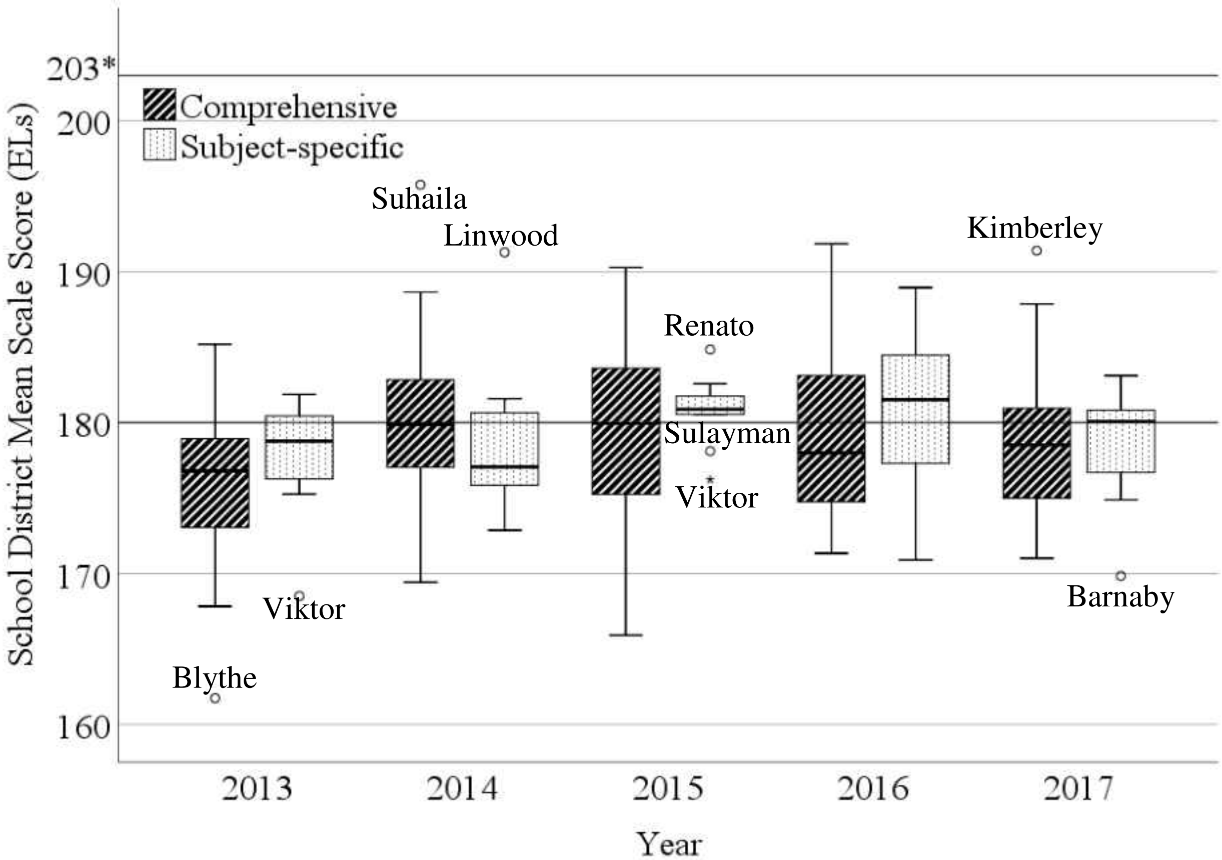

Figure 9. Boxplot of 2013–2017 school district mean scale scores for ELs, by school district group. Independent samples. *Minimum passing score (FLDOE, 2017o).

<u>Research Question 2</u>

Research Question 2: To what extent did 8th grade FCAT 2.0 Science/SSA school district mean scale scores for 2013, 2014, 2015, 2016, and 2017 differ between two groups of Florida school districts: those that offered comprehensive science courses and those that offered subject-specific science courses, when the school districts are matched by total student population size, low socioeconomic status (SES) student percentage, and English learner (EL) percentage?

To answer this question, paired samples *t*-tests were used to compare each year's 8th grade FCAT 2.0 Science/SSA school district mean scales scores for all students in paired samples of school districts that offered comprehensive science courses (comprehensive group) and school districts that offered subject-specific science courses (subject-specific group). The paired samples *t*-tests controlled for school district population, percentage of low SES students, and percentage of ELs by matching school districts that offered comprehensive science courses to school districts that offered subject-specific science courses by these demographic factors (see Chapter 3, Sampling Method section for the criteria used to match each school district pair). While this reduced the degrees of freedom in the samples, it ensured that the school district mean scale scores being analyzed were from comparable school districts, as verified by Pearson *r* correlations conducted for each demographic matching factor (Steinberg, 2011). These correlations are shown in Table 56. For each year's sample, skewness and kurtosis were examined to test the assumption of normality of distribution (Schmider et al., 2010), and Pitman-Morgan tests were used to test the assumption of homogeneity of variance in the paired samples (Gardner, 2001; Morgan, 1939; Pitman, 1939). All tests conducted were non-directional, $\alpha = .05$.

Table 56

School District Paired Samples Demographic Matching Correlations

Year	Demographic	*N* (school district pairs)	*r*	*p*
2013	Student population	18	.95	<.001
	Percentage of low SES students	18	.89	<.001
	Percentage of ELs	18	.88	<.001
2014	Student population	15	.93	<.001
	Percentage of low SES students	15	.90	<.001
	Percentage of ELs	15	.93	<.001
2015	Student population	14	.95	<.001
	Percentage of low SES students	14	.70	<.001
	Percentage of ELs	14	.95	<.001
2016	Student population	15	.98	<.001
	Percentage of low SES students	15	.86	<.001
	Percentage of ELs	15	.73	<.001
2017	Student population	17	.96	<.001
	Percentage of low SES students	17	.70	<.001
	Percentage of ELs	17	.89	<.001

2013 Paired Samples Tests of School District Mean Scale Scores for All Students

As shown in Table 57, the 2013 8th grade FCAT 2.0 Science/SSA school district mean scale score for all students was numerically higher for the comprehensive group than for the subject-specific group. The mean scale score distributions of both groups were sufficiently normal for the purposes of conducting a *t*-test, i.e., skew < |2.0| and kurtosis < |9.0| (Schmider et al., 2010). The assumption of homogeneity of variance was satisfied by Pitman-Morgan test, $F(16) = .25$, $p = .96$ (Gardner, 2001; Morgan, 1939; Pitman, 1939). As shown in Table 57, the paired samples *t*-test revealed no statistically significant difference in the 2013 FCAT 2.0

Science/SSA school district mean scale score for all students between the two school district groups in the paired sample.

Table 57

2013 Paired Samples Tests of School District Mean Scale Scores for All Students (*N* = 18 school district pairs)

School district group	Descriptive statistics							Paired samples test				
											95% CI	
	n	*M*	*Mdn*	*SD*	*SEM*	Skew	Kurtosis	*t*	*df*	*p*	*LL*	*UL*
C	18	200.79	201.94	5.51	1.30	-2.32	8.29	1.55	17	.14	-0.53	3.49
S	18	199.31	199.56	5.98	1.41	-1.47	4.65					

2014 Paired Samples Tests of School District Mean Scale Scores for All Students

As shown in Table 58, the 2014 FCAT 2.0 Science/SSA school district mean scale score for all students was numerically higher for the comprehensive group than for the subject-specific group. The mean scale score distributions for both groups were sufficiently normal for the purposes of conducting a *t*-test, i.e., skew < |2.0| and kurtosis < |9.0| (Schmider et al., 2010). The assumption of homogeneity of variance was satisfied by the Pitman-Morgan test, $F(13) = 0.18$, $p = .98$ (Gardner, 2001; Morgan, 1939; Pitman, 1939). As shown in Table 58, the paired samples *t*-test showed no statistically significant difference in the 2014 FCAT 2.0 Science/SSA school district mean scale score for all students between the two groups of school districts in the paired sample.

Table 58

2014 Paired Samples Tests of School District Mean Scale Scores for All Students (*N* = 15 school district pairs)

School district group	Descriptive statistics							Paired samples test				
											95% CI	
	n	*M*	*Mdn*	*SD*	*SEM*	Skew	Kurtosis	*t*	*df*	*p*	*LL*	*UL*
C	15	201.92	202.80	3.79	0.98	-0.03	-0.25	1.55	14	.14	-0.80	4.92
S	15	199.86	199.25	4.26	1.10	0.13	-0.69					

2015 Paired Samples Tests of School District Mean Scale Scores for All Students

As shown in Table 59, the 2015 FCAT 2.0 Science/SSA school district mean scale score for all students was numerically lower for the comprehensive group than for the subject-specific group. The mean scale score distributions for both groups were sufficiently normal for the purposes of conducting a *t*-test, i.e., skew < |2.0| and kurtosis < |9.0| (Schmider et al., 2010). The assumption of homogeneity of variance was satisfied by the Pitman-Morgan test, *F* (12) = 6.24, *p* = .15 (Gardner, 2001; Morgan, 1939; Pitman, 1939). Also shown in Table 59, the paired samples *t*-test revealed no statistically significant difference in the 2015 8th grade FCAT 2.0 Science/SSA school district mean scale score for all students between the two groups of school districts.

Table 59

2015 Paired Samples Tests of School District Mean Scale Scores for All Students (*N* =14 school district pairs)

School district group	Descriptive statistics							Paired samples test			95% CI	
	n	*M*	*Mdn*	*SD*	*SEM*	Skew	Kurtosis	*t*	*df*	*p*	*LL*	*UL*
C	14	200.46	201.00	4.93	1.32	-2.06	5.58	-0.19	13	.85	-3.37	2.82
S	14	200.73	201.20	2.53	0.68	0.13	-0.50					

2016 Paired Samples Tests of School District Mean Scale Scores for All Students

As shown in Table 60, the 2016 FCAT 2.0 Science/SSA school district mean scale score for all students was numerically higher for the comprehensive group than for the subject-specific group. The mean scale score distributions for both groups were sufficiently normal for the purposes of conducting a *t*-test, i.e., skew < |2.0| and kurtosis < |9.0| (Schmider et al., 2010). The assumption of homogeneity of variance was satisfied by the Pitman-Morgan test, *F* (13) = 0.93, *p* = .63 (Gardner, 2001; Morgan, 1939; Pitman, 1939). The paired samples *t*-test, shown in Table 60, revealed no statistically significant difference the 2016 8th grade FCAT 2.0 Science/SSA school district mean scale score for all students between the two groups of school districts.

Table 60

2016 Paired Samples Tests of School District Mean Scale Scores for All Students (*N* = 15 school district pairs)

School district group	Descriptive statistics							Paired samples test				
											95% CI	
	n	*M*	*Mdn*	*SD*	*SEM*	Skew	Kurtosis	*t*	*df*	*p*	*LL*	*UL*
C	15	202.14	202.00	3.91	1.01	-0.23	0.58	0.48	14	.64	-2.16	3.39
S	15	201.52	200.91	3.01	0.78	0.18	0.58					

2017 Paired Samples Tests of School District Mean Scale Scores for All Students

As shown in Table 61, the 2017 FCAT 2.0 Science/SSA school district mean scale score for all students was numerically higher for the comprehensive group than for the subject-specific group. The mean scale score distributions for both groups were sufficiently normal for the purposes of conducting a *t*-test, i.e., skew < |2.0| and kurtosis < |9.0| (Schmider et al., 2010). The assumption of homogeneity of variance was satisfied by the Pitman-Morgan test, F (15) = 4.17, p = .21 (Gardner, 2001; Morgan, 1939; Pitman, 1939). As shown in Table 61, the paired samples *t*-test revealed no statistically significant difference in the 2017 8th grade FCAT 2.0 Science/SSA school district mean scale score for all students between the two groups of school districts.

Table 61

2017 Paired Samples Tests of School District Mean Scale Scores for All Students (*N* = 17 school district pairs)

School district group	Descriptive statistics							Paired samples test			95% CI	
	n	*M*	*Mdn*	*SD*	*SEM*	Skew	Kurtosis	*t*	*df*	*p*	*LL*	*UL*
C	17	201.07	201.06	5.32	1.29	-0.32	-0.26	0.15	16	.88	-3.32	3.84
S	17	200.81	200.63	3.26	0.79	0.40	0.65					

Summary of Research Question 2 Paired Samples Tests

Research Question 2 used paired sample *t*-tests to analyze the differences in school district 8th grade FCAT 2.0 Science/SSA mean scale scores for all students between groups of school districts that offered comprehensive science courses and groups of school districts that offered subject-specific science courses. Each year's paired samples satisfied the tests for normality of distribution and homogeneity of variance. Each year's paired samples were matched with very strong correlations (Steinberg, 2011) for student population, percentage of low SES students, and percentage of ELs. As shown in

Figure 10, for four of the five years, the mean scale scores for all students were numerically higher for the comprehensive group than for the subject-specific group. This finding differed from that for the independent samples tests, for which the school district mean scale scores for all students were numerically higher for the subject-specific group each year. However, as with the independent samples tests, no statistically significant differences were found in the 2013, 2014, 2015, 2016, or 2017 8th grade FCAT 2.0 Science/SSA school district mean scale scores for all students between the two groups of school districts in the paired samples.

All mean scale scores were below the minimum passing scale score of 203. The descriptive statistics and paired samples *t*-test results for 2013 through 2017 for Research Question 2 are summarized in Table 62. The boxplot in Figure 11 shows that, as with the independent samples, the dispersions of the 2014, 2015, and 2016 paired sample school district mean scale scores were greater in the comprehensive group than the subject-specific group. Outliers consisted of the same school districts as the Research Question 1 independent samples, when those school districts were selected as part of the paired samples.

Table 62

2013–2017 Paired Samples Tests of School District Mean Scale Scores for All Students

							Paired samples *t*-tests				
	School district	Descriptive statistics								95% CI	
Year	group	*n*	*M*	*Mdn*	*SD*	*SEM*	*t*	*df*	*p*	*LL*	*UL*
2013	C	18	200.79	201.94	5.51	1.30	1.55	17	.14	-0.53	3.49
	S	18	199.31	199.56	5.98	1.41					
2014	C	15	201.92	202.80	3.79	0.98	1.55	14	.14	-0.80	4.92
	S	15	199.86	199.25	4.26	1.10					
2015	C	14	200.46	201.00	4.93	1.32	-0.19	13	.85	-3.37	2.82
	S	14	200.73	201.20	2.53	0.68					
2016	C	15	202.14	202.00	3.91	1.01	0.48	14	.64	-2.16	3.39
	S	15	201.52	200.91	3.01	0.78					
2017	C	17	201.07	201.06	5.32	1.29	0.15	16	.88	-3.32	3.84
	S	17	200.81	200.63	3.26	0.79					

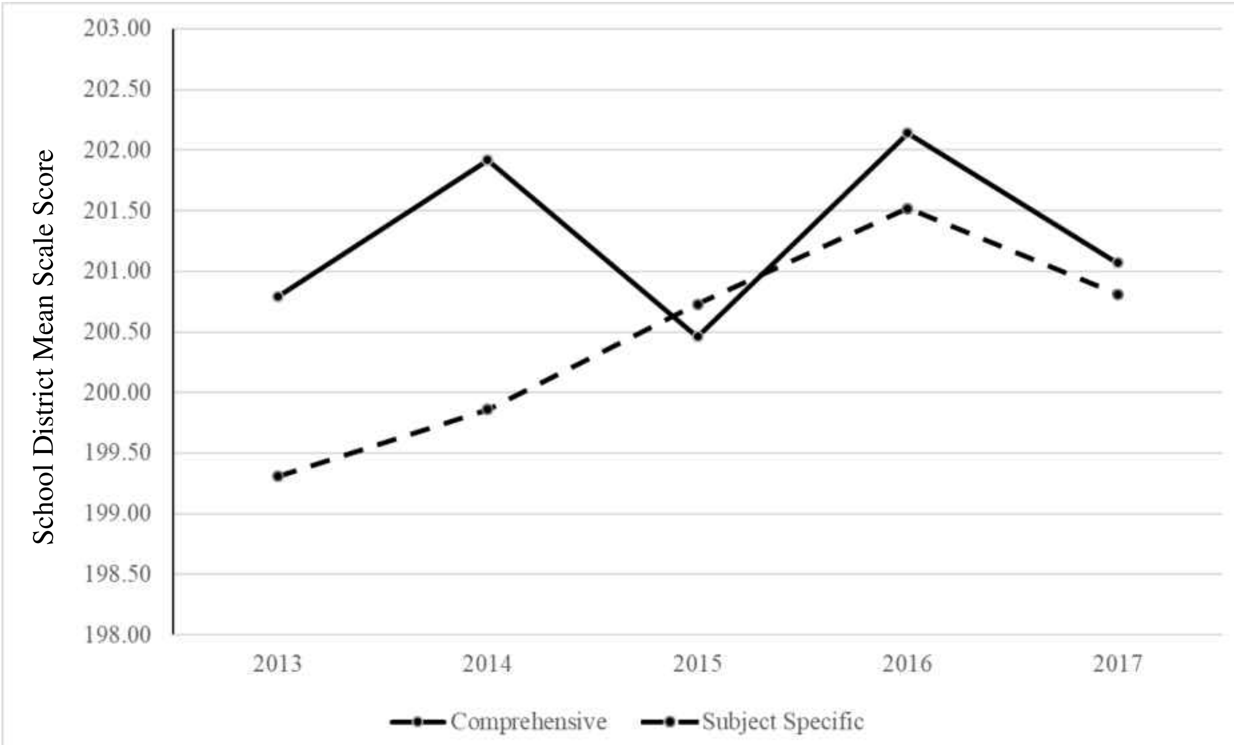

Figure 10. Paired samples 8th Grade FCAT 2.0 Science/SSA school district mean scale scores for all students, 2013–2017.

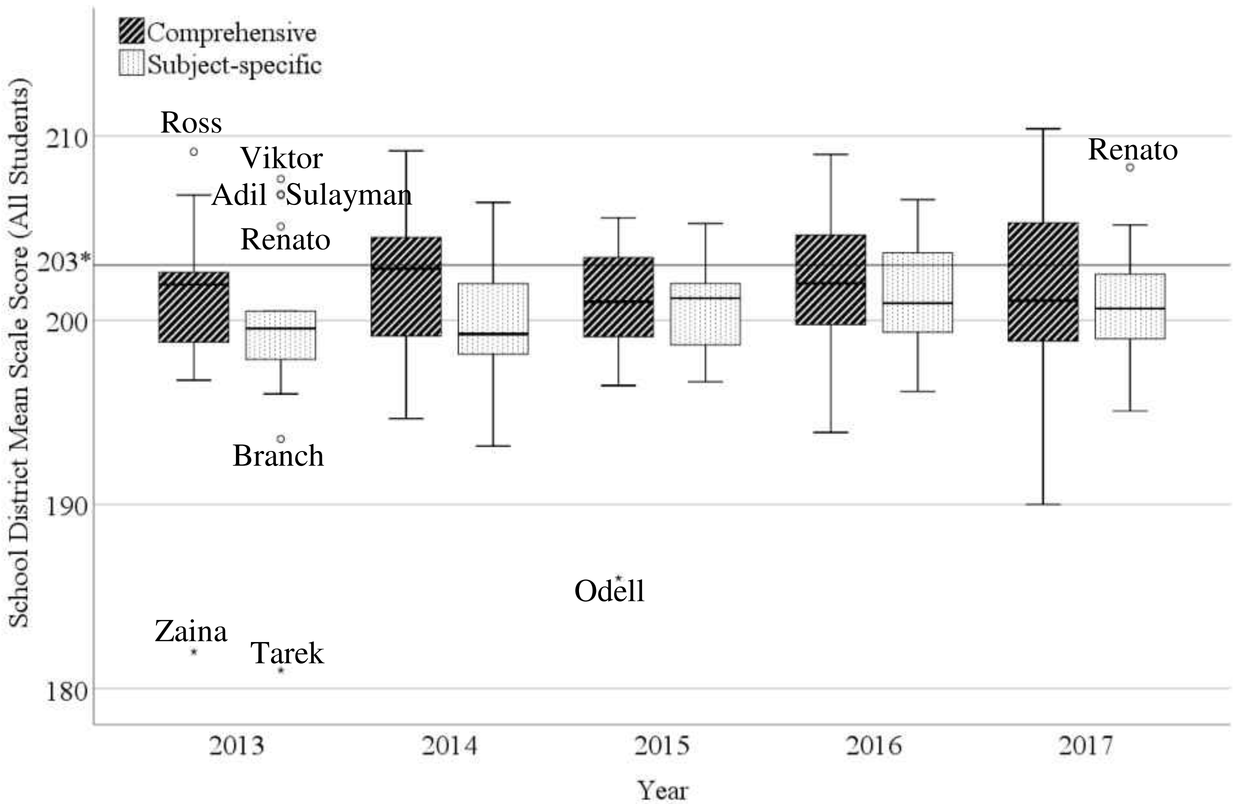

Figure 11. Boxplot of paired samples school district mean scale scores for all students, by school district group. *Minimum passing score (FLDOE, 2017o).

Research Question 3

Research Question 3: To what extent did 8th grade FCAT 2.0 Science/SSA school district mean scale scores for low SES students for 2013, 2014, 2015, 2016, and 2017 differ between two groups of Florida school districts: those that offered comprehensive science courses and those that offered subject-specific science courses, when the school districts are matched by total student population, low SES student percentage, and EL percentage?

To answer this question, paired samples *t*-tests were used to compare each year's 8th grade FCAT 2.0 Science/SSA mean scales scores for low SES students in the same paired samples used for Research Question 2. The paired samples *t*-tests controlled for school district population, percentage of low SES students, and percentage of ELs by matching school districts from each group by these demographic factors (see Chapter 3, Sampling Method section for the criteria used to match each school district pair). While this reduced the degrees of freedom in the samples, it ensured comparison of similar school districts. Skewness and kurtosis were examined to test the assumption of normality of distribution (Schmider et al., 2010), and Pitman-Morgan tests were used to test the assumption of homogeneity of variance of the differences in the paired samples (Gardner, 2001; Morgan, 1939; Pitman, 1939). All tests conducted were non-directional, $\alpha = .05$.

2013 Paired Samples Tests of School District Mean Scale Scores for Low SES Students

As shown in Table 63, the 2013 8^{th} grade FCAT 2.0 Science/SSA school district mean scale score for low SES students was numerically higher for the comprehensive group than for the subject-specific group. The mean scale score distributions of both groups were sufficiently normal for the purposes of conducting a *t*-test, i.e., skew < |2.0| and kurtosis < |9.0|; (Schmider et al., 2010). The assumption of homogeneity of variance was satisfied by Pitman-Morgan test, $F(16) = 3.13$, $p = .27$ (Gardner, 2001; Morgan, 1939; Pitman, 1939). Also shown in Table 63, the paired samples *t*-test revealed no statistically significant difference in the 2013 FCAT 2.0 Science/SSA school district mean scale score for low SES students between the two groups of school districts.

Table 63

2013 Paired Samples Tests of School District Mean Scale Scores for Low SES Students (*N* = 18 school district pairs)

School district group	Descriptive statistics							Paired samples test				
											95% CI	
	n	*M*	*Mdn*	*SD*	*SEM*	Skew	Kurtosis	*t*	*df*	*p*	*LL*	*UL*
C	18	189.85	187.14	6.95	1.64	0.21	-1.50	0.63	17	.54	-2.34	4.31
S	18	188.87	189.05	4.68	1.10	0.32	-0.23					

2014 Paired Samples Tests of School District Mean Scale Scores for Low SES Students

As shown in Table 64, the 2014 FCAT 2.0 Science/SSA school district mean scale score for low SES students was numerically higher for the comprehensive group than for the subject-specific group. The mean scale score distributions for both groups were sufficiently normal for the purposes of conducting a *t*-test, i.e., skew < |2.0| and kurtosis < |9.0|; (Schmider et al., 2010). The assumption of homogeneity of variance was satisfied by the Pitman-Morgan test, $F(13) = 0.36$, $p = .90$ (Gardner, 2001; Morgan, 1939; Pitman, 1939). Also shown in Table 64, the paired samples *t*-test revealed no statistically significant difference in the 2014 FCAT 2.0 Science/SSA school district mean scale score for low SES students between the two groups of school districts. The test of 2014 data approached significance, with a *p*-value just over the $\alpha = .05$ threshold.

Table 64

2014 Paired Samples Tests of School District Mean Scale Scores for Low SES Students (*N* = 15 school district pairs)

School district group	Descriptive statistics							Paired samples test				
											95% CI	
	n	*M*	*Mdn*	*SD*	*SEM*	Skew	Kurtosis	*t*	*df*	*p*	*LL*	*UL*
C	15	191.88	191.81	5.55	1.41	0.47	-0.89	1.83	14	.09	-0.49	6.10
S	15	189.08	186.95	4.75	1.23	1.27	1.56					

2015 Paired Samples Tests of School District Mean Scale Scores for Low SES Students

As shown in Table 65, the 2015 8th grade FCAT 2.0 Science/SSA school district mean scale score for low SES students was numerically lower for the comprehensive group than for the subject-specific group. The mean scale score distributions for both groups were sufficiently normal for the purposes of conducting a *t*-test, i.e., skew < |2.0| and kurtosis < |9.0|; (Schmider et al., 2010). The assumption of homogeneity of variance was satisfied by the Pitman-Morgan test, F (12) = 0.22, p = .64 (Gardner, 2001; Morgan, 1939; Pitman, 1939). Also shown in Table 65, the paired samples *t*-test revealed no statistically significant difference in the 2015 FCAT 2.0 Science/SSA school district mean scale score for low SES students between the two groups of school districts.

Table 65

2015 Paired Samples Tests of School District Mean Scale Scores for Low SES Students (*N* = 14 school district pairs)

School district group	Descriptive statistics							Paired samples test				
											95% CI	
	n	*M*	*Mdn*	*SD*	*SEM*	Skew	Kurtosis	*t*	*df*	*p*	*LL*	*UL*
C	14	190.41	188.63	4.99	1.33	0.59	-0.87	-0.02	13	.98	-1.46	1.43
S	14	190.43	188.57	4.66	1.24	0.56	-1.41					

2016 Paired Samples Tests of School District Mean Scale Scores for Low SES Students

The 2016 FCAT 2.0 Science/SSA school district mean scale score for low SES students was numerically higher for the comprehensive group than for the subject-specific group, as shown in Table 66. The mean scale score distributions were sufficiently normal for the purposes of conducting a *t*-test, i.e., skew < |2.0| and kurtosis < |9.0|; (Schmider et al., 2010). The assumption of homogeneity of variance was satisfied by the Pitman-Morgan test, $F(13) = 0.27$, $p = .61$ (Gardner, 2001; Morgan, 1939; Pitman, 1939). Also shown in Table 66, the paired samples *t*-test revealed no statistically significant difference in the 2016 FCAT 2.0 Science/SSA school district mean scale score for low SES students between the two groups of school districts.

Table 66

2016 Paired Samples Tests of School District Mean Scale Scores for Low SES Students (*N* = 15 school district pairs)

School district group	Descriptive statistics							Paired samples test				
											95% CI	
	n	*M*	*Mdn*	*SD*	*SEM*	Skew	Kurtosis	*t*	*df*	*p*	*LL*	*UL*
C	15	193.04	193.60	5.67	1.46	-0.16	-1.24	1.74	14	.10	-0.51	4.90
S	15	190.85	188.87	5.05	1.30	0.93	-0.05					

2017 Paired Samples Tests of School District Mean Scale Scores for Low SES Students

The 2017 FCAT 2.0 Science/SSA school district mean scale score for low SES students was numerically higher for the comprehensive group than for the subject-specific group, as shown in Table 67. The mean scale score distributions for both groups were sufficiently normal for the purposes of conducting a *t*-test (i.e., skew < |2.0| and kurtosis < |9.0|; (Schmider et al., 2010). The assumption of homogeneity of variance was satisfied by the Pitman-Morgan test, $F(15) = 0.10$, $p = .98$ (Gardner, 2001; Morgan, 1939; Pitman, 1939). Also shown in Table 67, the paired samples *t*-test revealed no statistically significant difference in the 2017 FCAT 2.0 Science/SSA school district mean scale score for low SES students between the two groups of school districts.

Table 67

2017 Paired Samples Tests of School District Mean Scale Scores for Low SES Students (*N* = 17 school district pairs)

School district group	Descriptive statistics							Paired samples test				
											95% CI	
	n	*M*	*Mdn*	*SD*	*SEM*	Skew	Kurtosis	*t*	*df*	*p*	*LL*	*UL*
C	17	190.13	188.86	5.60	1.36	0.87	-0.31	0.04	16	.97	-3.03	3.14
S	17	190.08	187.96	5.58	1.35	0.74	-1.10					

Summary of Research Question 3 Paired Samples Tests

Research Question 3 used paired samples *t*-tests to test the differences in 2013, 2014, 2015, 2016, and 2017 8th grade FCAT 2.0 Science/SSA school district mean scale scores for low SES students between two groups of school districts: those that offered comprehensive science courses (comprehensive group) and those that offered subject-specific science courses (subject-specific group). Each year's paired samples satisfied the tests for normality of distribution and homogeneity of variance. As shown in Figure 12, for four of the five years, the mean scale scores for low SES students were numerically higher for the comprehensive group than for the subject-specific group. This finding was similar to the independent samples tests, for which the school district mean scale scores for low SES students were numerically higher each year for the comprehensive group. However, as with the independent samples tests, the paired sample *t*-tests found no statistically significant differences in the 2013, 2014, 2015, 2016, or 2017 8th grade FCAT 2.0 Science/SSA school district mean scale scores for low SES students between the two groups of school districts in the paired samples. The test of 2014 data approached significance, with a *p*-value just over the $\alpha = .05$ threshold.

All school district mean scale scores were below the minimum passing scale score of 203. The descriptive statistics and paired samples *t*-test results for Research Question 3 are summarized in Table 68. The boxplot in Figure 13 shows that the dispersions of the school district mean scale scores for low SES students in the comprehensive group were greater than those of the subject-specific group. The boxplot also shows that the maximum school district mean scale score for low SES students in the paired samples did not reach the minimum passing score for either group of school districts in any year from 2013 to 2017.

Table 68

2013–2017 Paired Samples Tests of School District Mean Scale Scores for Low SES Students

		Descriptive statistics					Paired samples *t*-tests				
										95% CI	
Year	School district group	*n*	*M*	*Mdn*	*SD*	*SEM*	*t*	*df*	*p*	*LL*	*UL*
2013	C	18	189.85	187.14	6.95	1.64	0.63	17	.54	-2.34	4.31
	S	18	188.87	189.05	4.68	1.10					
2014	C	15	191.88	191.81	5.55	1.41	1.83	14	.089	-0.49	6.10
	S	15	189.08	186.95	4.75	1.23					
2015	C	14	190.41	188.63	4.99	1.33	-0.02	13	.98	-1.46	1.43
	S	14	190.43	188.57	4.66	1.24					
2016	C	15	193.04	193.60	5.67	1.46	1.74	14	.10	-0.51	4.90
	S	15	190.85	188.87	5.05	1.30					
2017	C	17	190.13	188.86	5.60	1.36	0.04	16	.97	-3.03	3.14
	S	17	190.08	187.96	5.58	1.35					

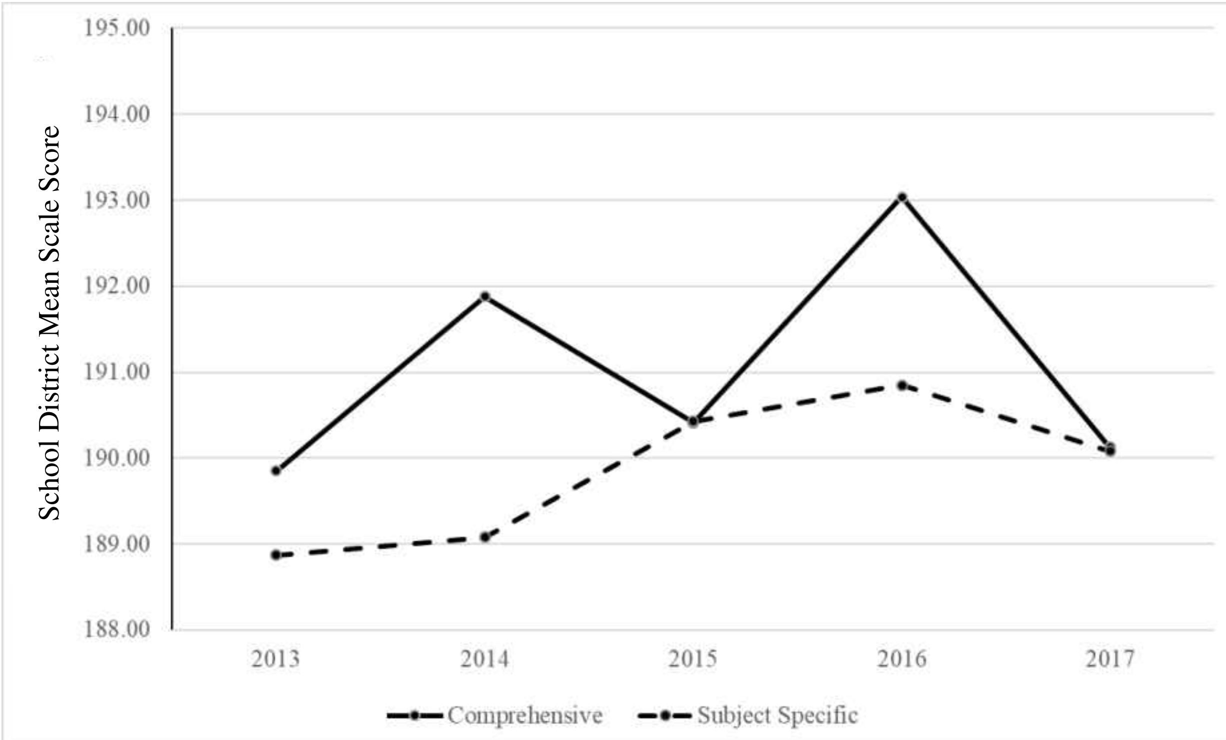

Figure 12. Paired samples 8th Grade FCAT 2.0 Science/SSA school district mean scale scores for low SES students.

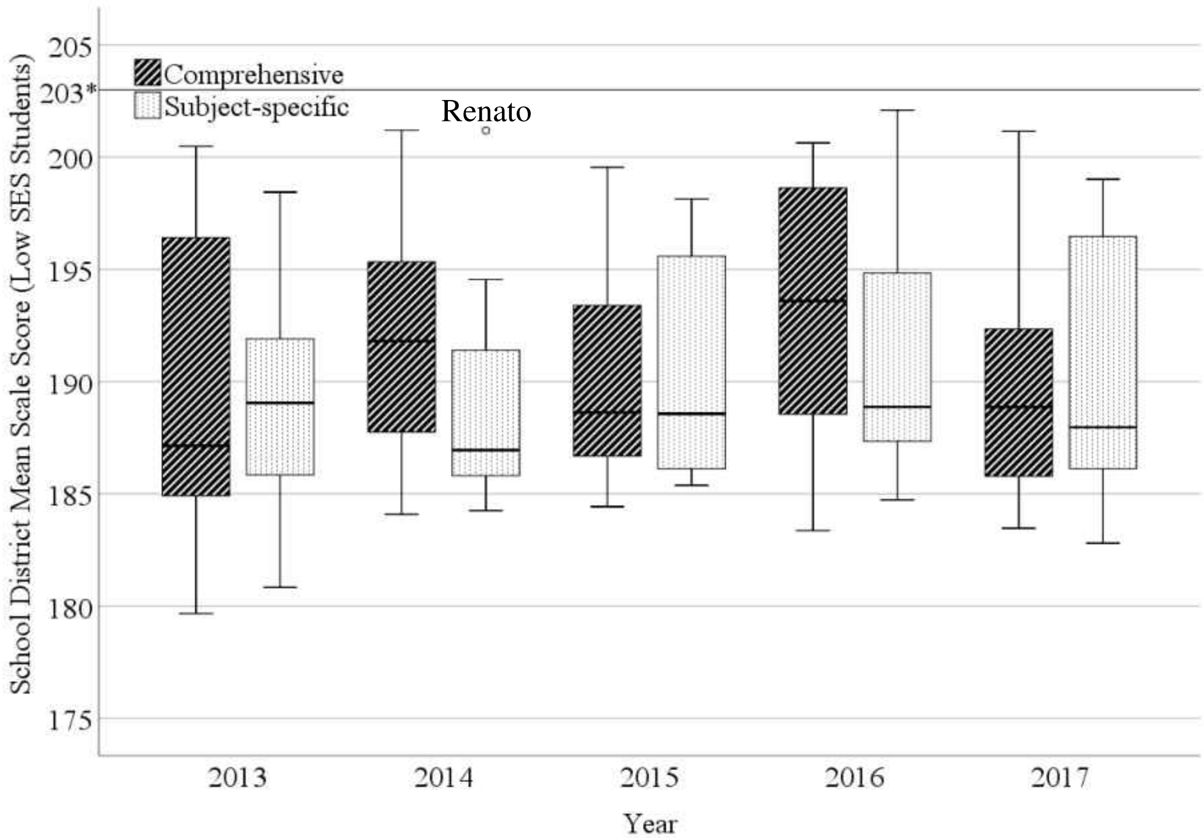

Figure 13. Boxplot of paired samples school district mean scale scores for low SES students, by school district group. *Minimum passing score (FLDOE, 2017o).

<u>Research Question 4</u>

Research Question 4: To what extent did 8th grade FCAT 2.0 Science/SSA school district mean scale scores for ELs for years 2013, 2014, 2015, 2016, and 2017 differ between two groups of Florida school districts: those that offered comprehensive science courses and those that offered subject-specific science courses, when the school districts are matched by total student population size, low SES student percentage, and EL percentage?

To answer this question, paired samples *t*-tests were used to compare each year's 8th grade FCAT 2.0 Science/SSA school district mean scales scores for ELs in paired school district samples. The paired samples *t*-tests controlled for school district population, percentage of low SES students, and percentage of ELs by matching school districts from each group by these demographic factors (see Chapter 3, Sampling Method section for the criteria used to match each school district pair). Due to suppression of mean scale score data for small school districts with fewer than 10 ELs assessed (FLDOE, 2017e), some school districts in the Research Questions 2 and 3 could not be used in the Research Question 4 samples, reducing the degrees of freedom. Skewness and kurtosis were examined to test the assumption of normality of distribution (Schmider et al., 2010), and Pitman-Morgan tests were used to test the assumption of homogeneity of variance of the differences in the paired samples (Gardner, 2001; Morgan, 1939; Pitman, 1939). All tests conducted were non-directional, $\alpha = .05$.

2013 Paired Samples Tests of School District Mean Scale Scores for ELs

The 2013 8th grade FCAT 2.0 Science/SSA school district mean scale score for ELs was numerically lower for the comprehensive group than for the subject-specific group, as shown in Table 69. The mean scale score distributions of both groups were sufficiently normal for the purposes of conducting a *t*-test, i.e., skew < |2.0| and kurtosis < |9.0| (Schmider et al., 2010). The assumption of homogeneity of variance was satisfied by Pitman-Morgan test, $F(7) = 1.96$, $p = .38$ (Gardner, 2001; Morgan, 1939; Pitman, 1939). As shown in Table 69, the paired samples *t*-test revealed no statistically significant difference in the 2013 FCAT 2.0 Science/SSA school district mean scale score for ELs between the two groups of school districts.

Table 69

2013 Paired Samples Tests of School District Mean Scale Scores for ELs (*N* = 9 school district pairs)

School district group	Descriptive statistics							Paired samples test				
											95% CI	
	n	*M*	*Mdn*	*SD*	*SEM*	Skew	Kurtosis	*t*	*df*	*p*	*LL*	*UL*
C	9	177.40	178.58	3.71	1.24	-1.27	0.67	-0.68	8	.52	-5.39	2.95
S	9	178.63	178.78	2.42	0.81	-0.02	-1.11					

2014 Paired Samples Tests of School District Mean Scale Scores for ELs

The 2014 FCAT 2.0 Science/SSA school district mean scale score for ELs was numerically higher for the comprehensive group than for the subject-specific group, as shown in Table 70. The mean scale score distributions of both groups were sufficiently normal for the purposes of conducting a *t*-test (i.e., skew < |2.0| and kurtosis < |9.0|; (Schmider et al., 2010). The assumption of homogeneity of variance was satisfied by the Pitman-Morgan test, $F(7) = 0.85$, $p = .64$ (Gardner, 2001; Morgan, 1939; Pitman, 1939). As shown in Table 70, the paired samples *t*-test revealed no statistically significant difference in the 2014 FCAT 2.0 Science/SSA school district mean scale score for ELs between the two groups of school district.

Table 70

2014 Paired Samples Tests of School District Mean Scale Scores for ELs (*N* = 9 school district pairs)

School district group	Descriptive statistics							Paired samples test				
											95% CI	
	n	*M*	*Mdn*	*SD*	*SEM*	Skew	Kurtosis	*t*	*df*	*p*	*LL*	*UL*
C	9	182.36	182.85	6.77	2.26	0.79	0.87	0.98	8	.36	-3.36	8.33
S	9	179.88	177.06	4.83	1.61	1.84	4.10					

2015 Paired Samples Tests of School District Mean Scale Scores for ELs

The 2015 FCAT 2.0 Science/SSA school district mean scale score for ELs was numerically higher for the comprehensive group than for the subject-specific group, as shown in Table 71. The mean scale score distributions of both groups were sufficiently normal for the purposes of conducting a *t*-test (i.e., skew < |2.0| and kurtosis < |9.0|; (Schmider et al., 2010). The assumption of homogeneity of variance was satisfied by the Pitman-Morgan test, $F(9) = 17.72$, $p < .05$ (Gardner, 2001; Morgan, 1939; Pitman, 1939). As shown in Table 71, the paired samples *t*-test revealed no statistically significant difference in the 2015 FCAT 2.0 Science/SSA school district mean scale score for ELs between the two groups of school districts.

Table 71

2015 Paired Samples Tests of School District Mean Scale Scores for ELs (N = 9 school district pairs)

School district group	Descriptive statistics							Paired samples test				
											95% CI	
	n	*M*	*Mdn*	*SD*	*SEM*	Skew	Kurtosis	*t*	*df*	*p*	*LL*	*UL*
C	9	183.04	182.57	4.42	1.47	0.36	-0.85	1.64	8	.14	-0.72	4.28
S	9	181.26	180.89	1.80	0.60	0.46	2.28					

2016 Paired Samples Tests of School District Mean Scale Scores for ELs

As shown in Table 72, the 2016 FCAT 2.0 Science/SSA school district mean scale score for ELs was numerically lower for the comprehensive group than for the subject-specific group. The mean scale score distributions of both groups were sufficiently normal for the purposes of conducting a *t*-test (i.e., skew < |2.0| and kurtosis < |9.0|; (Schmider et al., 2010). The assumption of homogeneity of variance was satisfied by the Pitman-Morgan test, $F(7) < .01$, p = 1.00 (Gardner, 2001; Morgan, 1939; Pitman, 1939). As shown in Table 72, the paired samples *t*-test revealed no statistically significant difference in the 2016 FCAT 2.0 Science/SSA school district mean scale score for ELs between the two groups of school districts. The test of 2016 data approached significance, with a *p*-value just over the $\alpha = .05$ threshold.

Table 72

2016 Paired Samples Tests of School District Mean Scale Scores for ELs (*N* = 9 school district pairs)

School district group	Descriptive statistics							Paired samples test				
											95% CI	
	n	*M*	*Mdn*	*SD*	*SEM*	Skew	Kurtosis	*t*	*df*	*p*	*LL*	*UL*
C	9	184.14	186.34	5.29	1.76	-0.24	-1.49	2.08	8	.071	-0.26	4.97
S	9	181.78	181.53	5.31	1.77	-1.05	1.41					

2017 Paired Samples Tests of School District Mean Scale Scores for ELs

The 2017 FCAT 2.0 Science/SSA school district mean scale score for ELs was numerically higher for the comprehensive group than for the subject-specific group, as shown in Table 73. The mean scale score distributions of both groups were sufficiently normal for the purposes of conducting a *t*-test, i.e., skew < |2.0| and kurtosis < |9.0| (Schmider et al., 2010). The assumption of homogeneity of variance was satisfied by the Pitman-Morgan test, *F* (9) = 1.98, *p* = .38 (Gardner, 2001; Morgan, 1939; Pitman, 1939). The paired samples t-test, also shown in Table 73, revealed no statistically significant difference in the 2017 FCAT 2.0 Science/SSA school district mean scale score for ELs between the two groups of school districts.

Table 73

2017 Paired Samples Tests of School District Mean Scale Scores for ELs (*N* = 11 school district pairs)

School district group	Descriptive statistics							Paired samples test				
											95% CI	
	n	*M*	*Mdn*	*SD*	*SEM*	Skew	Kurtosis	*t*	*df*	*p*	*LL*	*UL*
C	11	180.66	180.85	5.84	1.76	0.59	-0.60	1.35	10	.21	-1.39	5.67
S	11	178.52	180.08	3.91	1.18	-1.03	1.14					

Summary of Research Question 4 Paired Samples Tests

Research Question 4 used paired samples *t*-tests to test the differences in 2013, 2014, 2015, 2016 and 2017 8th grade FCAT 2.0 Science/SSA school district mean scale scores for ELs between two groups of school districts: those that offered comprehensive science courses and those that offered subject-specific science courses. Each year's paired samples satisfied the tests for normality of distribution and homogeneity of variance. School districts that offered different types of science courses were closely matched by student population, percentage of low SES students, and percentage of ELs. Sample sizes for each year's paired samples were small due to suppression of data for school districts with fewer than ten ELs assessed (FLDOE, 2017e).

As shown in Figure 14, for four of the five years, the mean scale scores for ELs were numerically higher for the comprehensive group than for the subject-specific group. This finding differs from that of the independent samples tests, for which the school district mean scale scores for ELs were numerically higher for the subject specific group in three of the five years. However, as with the independent samples tests, no statistically significant differences were found in the 2013, 2014, 2015, 2016, or 2017 school district mean scale scores for ELs between the two groups in the paired samples. As shown in Table 74, the test of 2016 data approached significance, with a *p*-value just over the $\alpha = .05$ threshold. Each year, all school district mean scale scores for ELs were below the minimum passing scale score of 203, regardless of the type of science course offered.

The boxplot in Figure 15 shows that, except for 2016, the dispersion of the comprehensive group school district mean scale scores for ELs was greater each year than that of the subject-specific group. The boxplot also shows the maximum school district mean scale score for ELs in all school districts, regardless of the type of science course offered, did not reach the minimum passing score in any year from 2013 to 2017.

Table 74

2013–2017 Paired Samples Tests of School District Mean Scale Scores for ELs

							Paired samples *t*-tests				
	School district	Descriptive statistics								95% CI	
Year	group	*n*	*M*	*Mdn*	*SD*	*SEM*	*t*	*df*	*p*	*LL*	*UL*
2013	C	9	177.40	178.58	3.71	1.24	-0.68	8	.52	-5.39	2.95
	S	9	178.63	178.78	2.42	0.81					
2014	C	9	182.36	182.85	6.77	2.26	0.98	8	.36	-3.36	8.33
	S	9	179.88	177.06	4.83	1.61					
2015	C	9	183.04	182.57	4.42	1.47	1.64	8	.14	-0.72	4.28
	S	9	181.26	180.89	1.80	0.60					
2016	C	9	184.14	186.34	5.29	1.76	2.08	8	.071	-0.26	4.97
	S	9	181.78	181.53	5.31	1.77					
2017	C	11	180.66	180.85	5.84	1.76	1.35	10	.21	-1.39	5.67
	S	11	178.52	180.08	3.91	1.18					

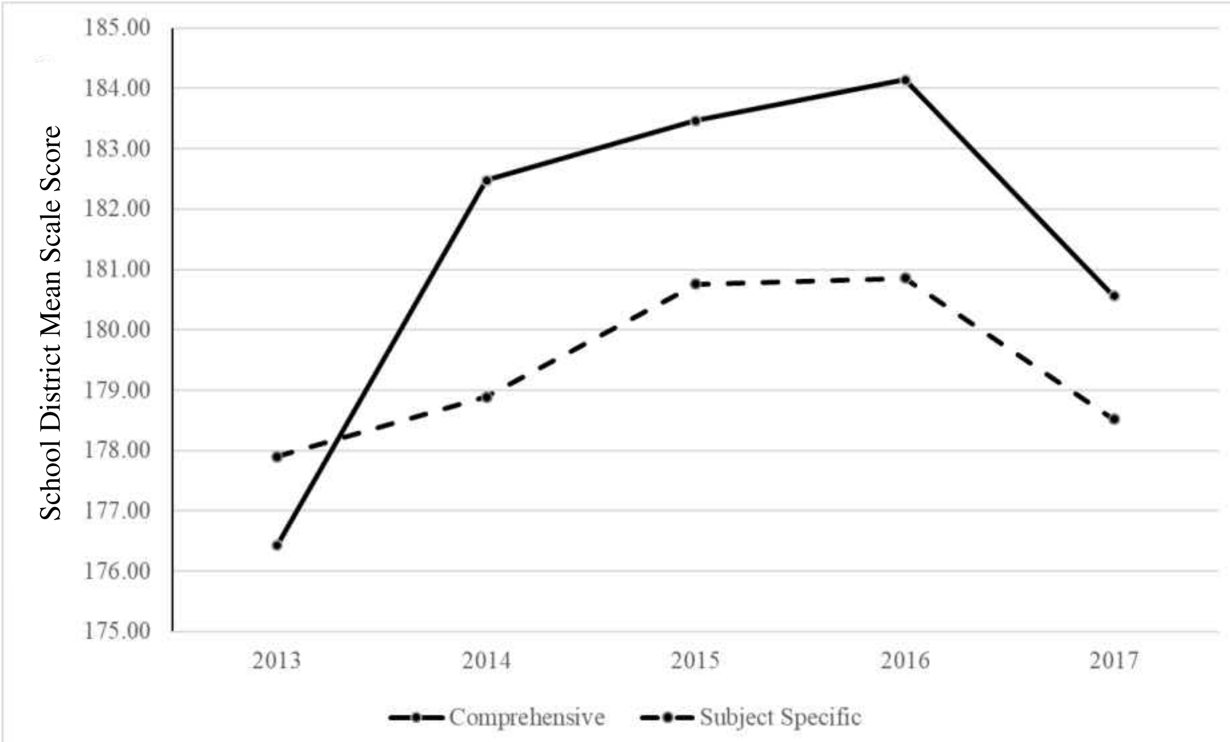

Figure 14. Paired samples 8th Grade FCAT 2.0 Science/SSA school district mean scale scores for ELs.

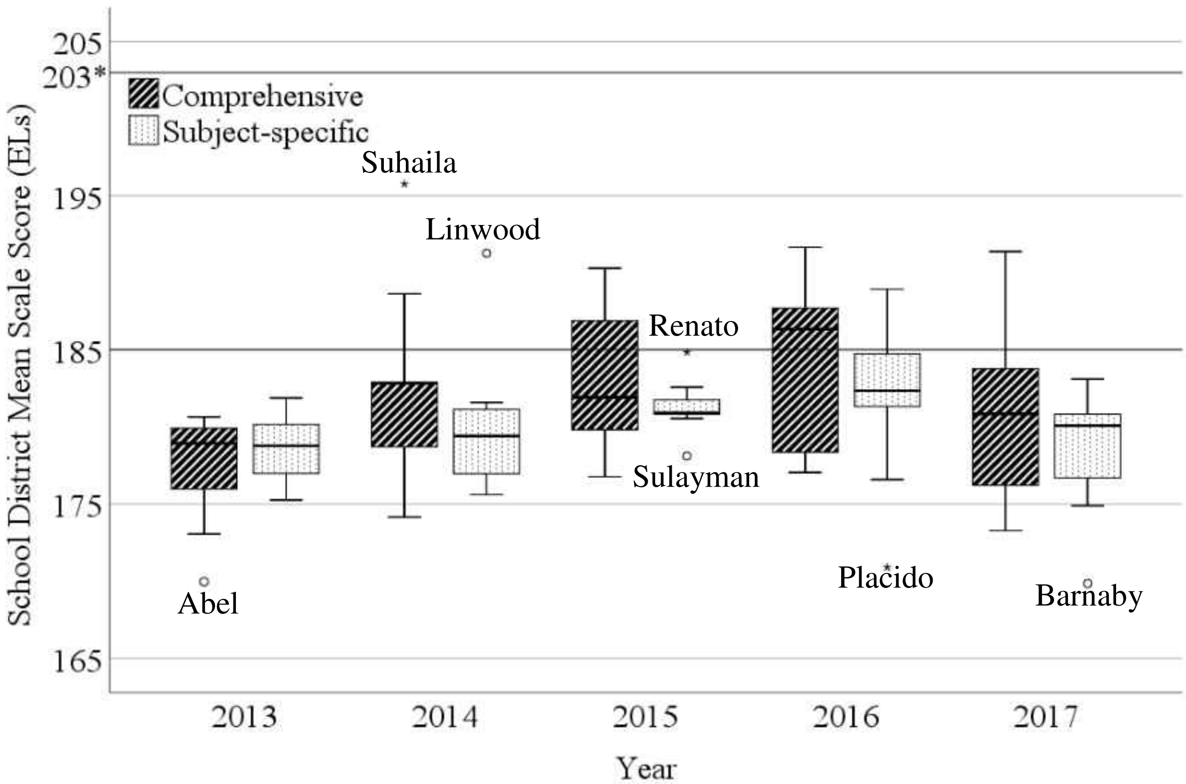

Figure 15. Boxplot of paired samples school district mean scale scores for ELs, by school district group. *Minimum passing score (FLDOE, 2017o).

Additional Analyses

The tests conducted for the four research questions revealed no significant differences in the 2013, 2014, 2015, 2016, or 2017 8^{th} grade FCAT 2.0 Science/SSA school district mean scale scores for three groups of students: all students, low SES students, and ELs; between the two groups of school districts. While each sample distribution met the *t*-test assumptions of normal distribution and equal variance (or the degrees of freedom adjusted), outliers and differences in dispersion between the two groups were evident. To check the validity of the *t*-tests, and to provide points for comparison in interpreting the findings, additional analyses were conducted to answer these seven questions:

1. To what extent did the 2013, 2014, 2015, 2016, and 2017 8th grade FCAT 2.0 Science/SSA school-level (as opposed to school district-level) mean scale scores for all students differ between two groups of Florida schools: those that offered comprehensive science courses and those that offered subject-specific science courses?

2. Do non-parametric tests of the Research Questions 1, 2, 3, and 4 samples indicate any statistically significant differences in the 2013, 2014, 2015, 2016, or 2017 8th grade FCAT 2.0 Science/SSA school district mean scale scores for the three groups of students between the two groups of school districts?

3. To what extent did the 8th grade FCAT 2.0 Science/SSA school district mean scale scores for three groups of students change between 2013 and 2017 for each group of school districts?

4. To what extent did the 2013, 2014, 2015, 2016, or 2017 8th grade FCAT 2.0 Science/SSA school district pass rates (percentage of students who achieved Level 3 or higher) for all students differ between the two school district groups?

5. To what extent did the 2013–2017 8th grade FCAT 2.0 Science/SSA school district mean raw scores for all students, by subject area (nature of science, physical science, life science, and Earth/space science), differ between the two groups of school districts?

6. To what extent was school district student population correlated with the 2013–2017 8th grade FCAT 2.0 Science/SSA school district mean scale scores for all students, low SES students, regardless of the of science course type offered?

7. Did changing the type of course offering have any general impact on school district mean scale scores for all students?

Analyses of School-level Mean Scale Scores

To what extent did the 2013, 2014, 2015, 2016, and 2017 8th grade FCAT 2.0 Science/SSA school-level (as opposed to school district-level) mean scale scores for all students, differ between two groups of Florida schools: those that offered comprehensive science courses and those that offered subject-specific science courses? The data used in the independent and paired samples tests for Research Questions 1–4 were two levels removed from the individual scores of the target population. The low degrees of freedom reduced the statistical power of the tests conducted on school district mean scale scores, and increased the chance of Type II error (Steinberg, 2011). As a means of validating the Research Question 1 results, additional independent samples *t*-tests were conducted to test the differences in school-level mean scale scores for all students between two groups of schools: those that offered comprehensive science courses (comprehensive group), and those that offered subject-specific science courses (subject-specific group). School-level mean scale scores for low SES students and ELs were not consistently available due to masking of data for schools with fewer than ten students assessed in these categories (FLDOE, 2017e).

As shown in Table 75, each year, the school mean scale scores for all students were numerically higher for the subject-specific group than for the comprehensive group. This finding is the same as that of the independent samples tests of school district mean scale scores for all students. As with the independent and paired samples tests of school district data, no statistically significant differences were found in the 2013, 2014, 2015, 2016, or 2017 8th grade FCAT 2.0 Science/SSA school-level mean scale scores for all students between the two groups of schools. All school-level mean scale scores were below the minimum passing scale score of 203. The boxplot in Figure 16 shows that, except for 2016, the dispersion of school mean scale scores for all students in the comprehensive group was greater than that of the subject-specific group. Outliers were evident in both groups of schools. The tests of school-level mean scale scores for all students support the validity of the Research Questions 1 and 2 tests of school district mean scale scores for all students. Detailed results of the school-level mean scale score analyses, by year, are at Appendix E.

Table 75

2013–2017 Independent Samples Tests of School Mean Scale Scores for All Students

		Descriptive statistics					Independent samples *t*-tests				
										95% CI	
Year	School group	*N*	*M*	*Mdn*	*SD*	*SEM*	*t*	*df*	*p*	*LL*	*UL*
2013	C	401	199.30	200.00	8.53	0.43	-0.85	571	.40	-2.15	0.85
	S	172	199.95	200.00	8.00	0.61					
2014	C	410	199.90	200.00	8.58	0.42	-0.65	578	.52	-1.99	1.00
	S	170	200.39	200.00	7.71	0.59					
2015	C	438	199.26	199.00	8.77	0.42	-1.55	593	.12	-2.88	0.34
	S	157	200.54	201.00	8.95	0.71					
2016	C	424	199.28	199.00	8.98	0.44	-1.34	579	.18	-2.74	0.52
	S	157	200.66	200.00	8.57	0.68					
2017	C	405	199.28	199.00	9.46	0.47	-1.25	568	.21	-3.70	0.82
	S	165	200.24	200.00	8.73	0.68					

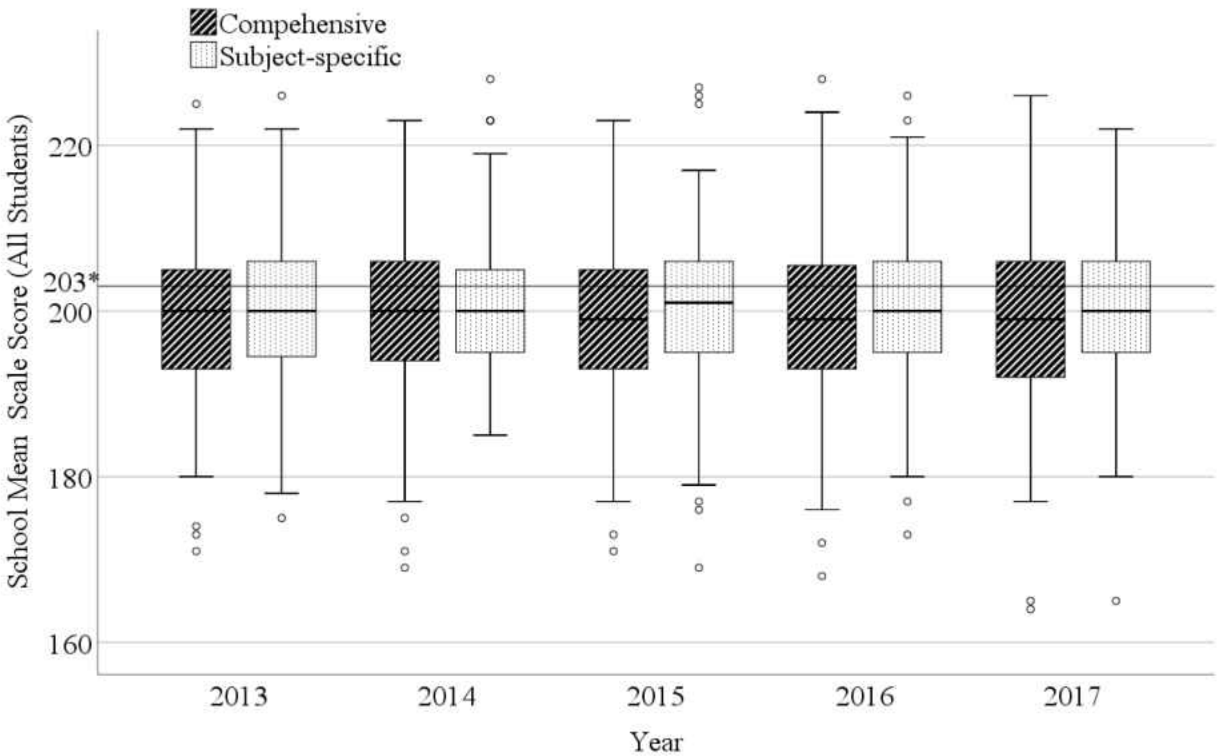

Figure 16. Boxplot of school mean scale scores for all students, by school district group.

Non-parametric Tests

Do non-parametric tests of the Research Questions 1, 2, 3, and 4 samples indicate any statistically significant differences in the 2013, 2014, 2015, 2016, or 2017 8th grade FCAT 2.0 Science/SSA school district mean scale scores for the three groups of students between the two groups of school districts? All samples, both independent and paired, used for Research Questions 1–4, met the requirements for normality of distribution for the purposes of conducting a *t*-test, i.e., skew < |2.0| and kurtosis < |9.0| (Schmider et al., 2010). Additionally, all samples either satisfied the assumption of equality of variances for conducting a *t*-test (Levene's or Pitman-Morgan tests) or were adjusted in degrees of freedom to account for unequal variances. The *t*-tests and the boxplots of distributions indicated that school districts that offered comprehensive science courses had greater dispersion of school district mean scale scores than those that offered subject-specific science courses. The boxplots of school district and school-level mean scale scores revealed the presence of outliers and extreme outliers in many samples.

Because of these distribution anomalies and small sample sizes, particularly for the paired samples *t*-tests of school district mean scale scores for ELs, non-parametric tests (Mann-Whitney *U* tests and Wilcoxon signed rank tests) were conducted to check the results of the Research Questions 1, 2, 3, and 4 *t*-tests. All tests were non-directional, $\alpha = .05$. All tests were conducted such that M_1 was the school district mean scale score for the comprehensive group, and M_2 was the school district mean scale score for the subject-specific group. Thus, a positive *Z*-score indicated that the sum of the ranks of mean scale scores for the comprehensive group was numerically lower than that of the subject-specific group, and vice versa.

For the independent samples used in the Research Question 1 *t*-tests, Mann-Whitney *U*-tests were conducted. The Mann-Whitney *U*-test, which can detect differences in shape, spread, and medians, is even more robust than the *t*-test to small sample size, non-normal distribution, and unequal variance (Hart, 2001). It often is used as an alternative when *t*-test assumptions are violated (Hart, 2001).

The results of the Mann-Whitney *U*-tests, shown in Table 76, corroborated the results of Research Question 1 independent samples *t*-tests. There were no statistically significant differences in the 2013, 2014, 2015, 2016, or 2017 8th grade FCAT 2.0 Science/SSA school district mean scale scores for all students, low SES students, or ELs, between the two groups of school districts. Like the Research Question 1 independent samples *t*-test, the Mann-Whitney *U*-test of 2014 mean scale scores for low SES students approached significance, with a *p*-value just over the α = .05 threshold, and the median school district mean scale score for low SES students numerically higher for the comprehensive group than for the subject-specific group. Unlike the independent samples *t*-test of 2015, 2016, and 2017 mean scale scores for all students, the Mann-Whitney tests for these years did not approach significance.

Table 76

Mann-Whitney U Tests of Independent Samples of School District Mean Scale Scores

		Student group		
Year	Statistic	All students	Low SES students	ELs
	Mdn (C*)	199.00	190.07	178.78
	Mdn (S)	199.56	189.05	176.80
2013	*U*	425.00	392.00	82.00
	Z	-0.23	-0.69	-1.49
	p	.82	.49	.14
	Mdn (C)	200.11	192.03	179.88
	Mdn (S)	199.25	186.95	177.06
2014	*U*	389.00	261.00	100.00
	Z	-0.02	-1.94	-0.80
	p	.99	.052	.42
	Mdn (C)	199.90	190.79	179.94
	Mdn (S)	201.20	188.57	180.89
2015	*U*	285.50	349.00	124.00
	Z	-1.32	-.34	-.38
	p	.19	.73	.71
	Mdn (C)	199.43	190.18	178.00
	Mdn (S)	200.92	188.87	181.53
2016	*U*	295.50	382.00	132.50
	Z	-1.42	-0.12	-0.82
	p	.16	.90	.42
	Mdn (C)	199.00	189.06	178.53
	Mdn (S)	200.63	187.96	180.08
2017	*U*	336.00	405.00	128.00
	Z	-1.18	-0.17	-0.33
	p	.24	.87	.74

Note. *C = comprehensive group, S = subject-specific group.

For the paired samples used in the Research Questions 2, 3, and 4, Wilcoxon signed rank tests, an alternative to paired samples *t*-tests (Hart, 2001), were conducted. The Wilcoxon signed rank tests corroborated the results of the paired samples *t*-tests conducted for Research Questions 2, 3, and 4. The results for these tests are shown in Table 77. No statistically significant differences were found in the 2013, 2014, 2015, 2016, or 2017 school district FCAT 2.0 Science/SSA mean scale scores for all students, low SES students, or ELs, between the two groups of school districts in the paired samples. Like the Research Question 4 paired samples *t*-test, the Wilcoxon signed ranks test of 2016 school district mean scale scores for ELs approached significance, with a *p*-value just over the $\alpha = .05$ threshold. The median school district mean scale score was numerically higher for the comprehensive group than for the subject-specific group. No other samples approached significance.

Table 77

Wilcoxon Signed Ranks Tests of Paired Samples of School District Mean Scale Scores

		Student group		
Year	Statistic	All students	Low SES students	ELs
	Mdn (C*)	201.94	187.14	178.58
	Mdn (S)	199.56	189.05	178.78
2013	*N*	18	18	9
	Z	-1.29	-0.85	-0.18
	p	.20	.40	.86
	Mdn (C)	202.80	191.81	182.85
	Mdn (S)	199.25	186.95	177.06
2014	*N*	15	15	9
	Z	-1.02	-1.48	-0.77
	p	.31	.14	.44
	Mdn (C)	201.00	188.63	182.57
	Mdn (S)	201.20	188.57	180.89
2015	*N*	14	14	9
	Z	-0.53	-0.03	-1.48
	p	.59	.98	.14
	Mdn (C)	202.00	193.60	186.34
	Mdn (S)	200.91	188.87	181.53
2016	*N*	15	15	9
	Z	-0.34	-1.31	-1.84
	p	.73	.19	.066
	Mdn (C)	201.06	188.86	180.85
	Mdn (S)	200.63	187.96	180.08
2017	*N*	17	17	11
	Z	-0.31	-0.26	-1.16
	p	.76	.80	.25

Note. *C = comprehensive group, S = subject-specific group.

Changes in School District Mean Scale Scores from 2013 to 2017

To what extent did the 8th grade FCAT 2.0 Science/SSA school district mean scale scores for the three groups of students change, from 2013 to 2017, for each group of school districts?

The findings of no significant differences in each year's mean scale scores for the three groups of students, between the two groups of school districts, may conceal significant longer-term changes in school district mean scale scores over the five-year period analyzed. Independent samples *t*-tests were conducted to analyze changes in the school district mean scale scores from 2013 to 2017, for the three groups of students, between the two groups of school districts. All tests were non-directional, $\alpha = .05$. All tests were conducted such that the 2013 mean scale scores were M_1, and 2017, M_2. Thus, negative t values indicate an increase in mean scale scores from 2013 to 2017, and positive t values indicate a decrease. Effect size (Cohen's d) was calculated for statistically significant changes.

As shown in Table 78, there were no statistically significant changes from 2013 to 2017 in the 8th grade FCAT 2.0 Science/SSA school district mean scale scores for all students or low SES students in either group of school districts. Also shown in Table 78, there was no statistically significant change from 2013 to 2017 in the 8th grade FCAT 2.0 Science/SSA school district mean scale scores for ELs in the subject-specific group of school districts. The only statistically significant change found was between the 2013 and 2017 school district mean scale scores for ELs in the comprehensive group of school districts. The 8th grade FCAT 2.0 Science/SSA school district mean scale scores for ELs in the comprehensive group increased

3.19 scale points from 2013 to 2017. The independent samples *t*-test showed this was a statistically significant difference, with a medium effect size, *d* = -0.49 (Cohen, 1988).

Table 78

Changes in School District Mean Scale Scores from 2013 to 2017, by School District Group

			Descriptive statistics					Independent samples *t*-tests				
											95% CI	
Student group	Year	School district group	*n*	*M*	*Mdn*	*SD*	*SEM*	*t*	*df*	*p*	*LL*	*UL*
All students	2013	C	49	198.94	199.00	5.84	0.83	0.27	96	.79	-2.20	2.90
	2017	C	49	198.59	199.00	6.85	0.98					
	2013	S	18	199.31	199.56	5.98	1.41	-0.92	33	.37	-4.85	1.84
	2017	S	17	200.81	200.63	3.26	0.79					
Low SES students	2013	C	49	190.67	190.07	7.42	1.06	-0.09	96	.93	-2.92	2.66
	2017	C	49	190.78	189.06	6.44	0.92					
	2013	S	18	188.87	189.05	4.69	1.10	-0.71	33	.49	-4.76	2.31
	2017	S	17	190.08	187.96	5.58	1.35					
ELs	2013	C	22	175.66	178.78	5.16	1.10	-2.14	45	.04*	-6.20	-0.18
	2017	C	25	178.85	178.53	5.06	1.01					
	2013	S	11	177.90	176.80	3.85	1.16	-0.37	20	.71	-4.07	2.83
	2017	S	11	178.52	180.08	3.91	1.18					

Note. *Effect size *d* = -0.49.

A boxplot of distributions of school district mean scale scores for all students, low SES students, and ELs, for each group of school districts (shown in Figure 17), was used to examine these differences more closely. The boxes for the 2013 and 2017 EL mean scale scores in the comprehensive group (which had a statistically significant change according to the *t*-test) revealed the presence of a low mean scale score outlier in 2013, and a high mean scale score outlier in 2017. The median of the 2013 comprehensive group mean scale score (*Mdn* = 178.78) was almost identical to the median of the 2017 comprehensive group mean scale score (*Mdn* = 178.53). After removing these outliers from the samples, the comprehensive group mean scale score for ELs students remained numerically higher in 2017 ($n = 24$, $M = 178.70$, $SD = 5.11$) than in 2013 ($n = 21$, $M = 176.32$, $SD = 4.22$). With the outliers removed, the *t*-test results, $t(43) = -1.69$, $p = .10$, showed the difference was not statistically significant. This shows that the significant change result of the initial in the *t*-test was caused by the non-offsetting outliers. Detailed results of the comparisons of 2013 to 2017 school district mean scale scores are at Appendix F.

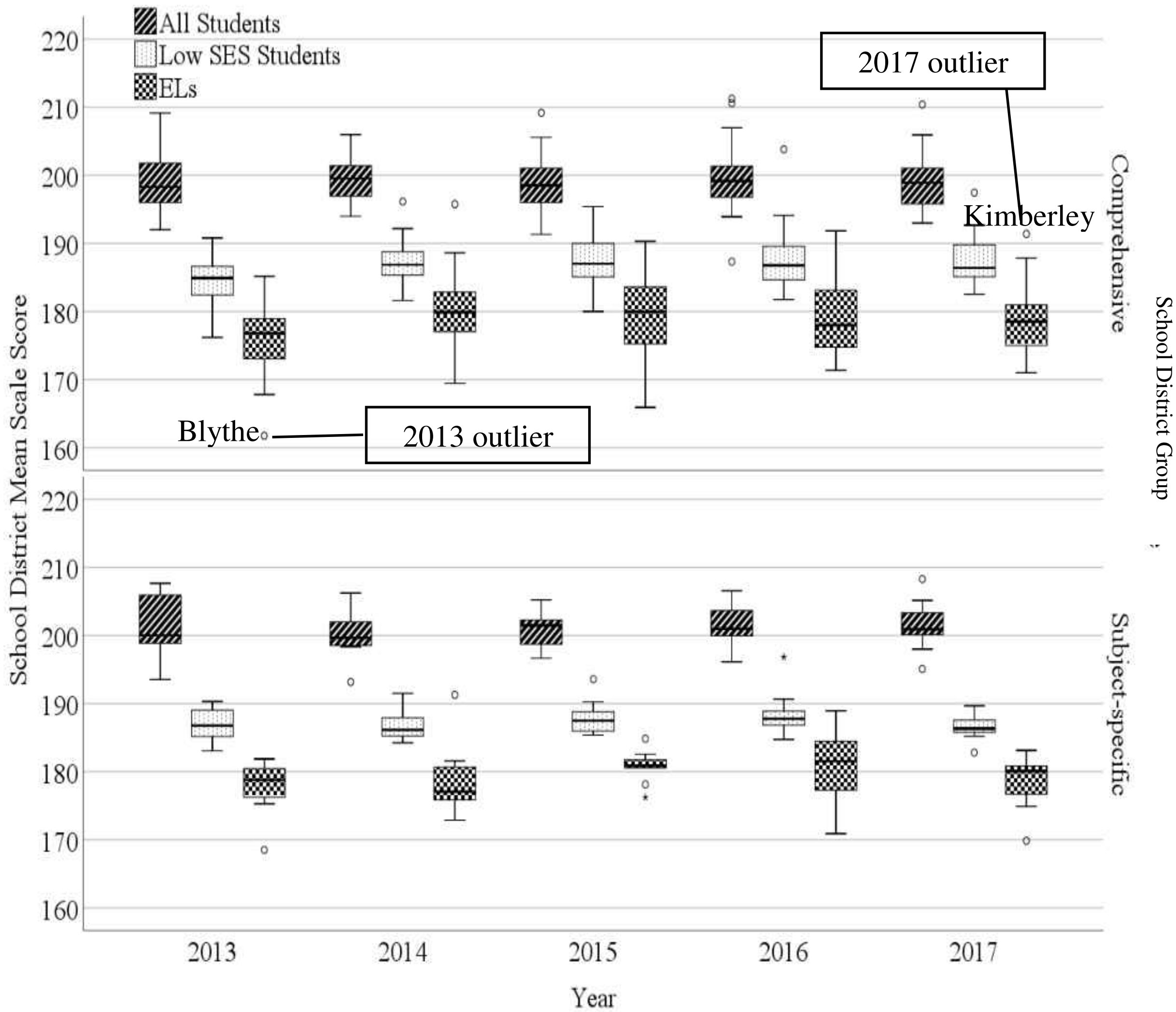

Figure 17. Boxplot of 2013–2017 school district mean scale score distributions, by school district group.

Analyses of School District Mean Pass Rates

To what extent did the 8th grade FCAT 2.0 Science/SSA school district mean pass rates (percentage of students who achieved Level 3 or higher) for all students differ between the two groups of school districts? As shown in Figure 18, each year from 2013 to 2017, the school district mean pass rate for all students in the subject-specific group increased, while that of the comprehensive group decreased. Because of the mean's susceptibility to the influence of outliers, and the possibility of offsetting changes on either side of the mean and median (Steinberg, 2011), analyses of mean scale scores may have masked significant changes in the school district mean pass rates.

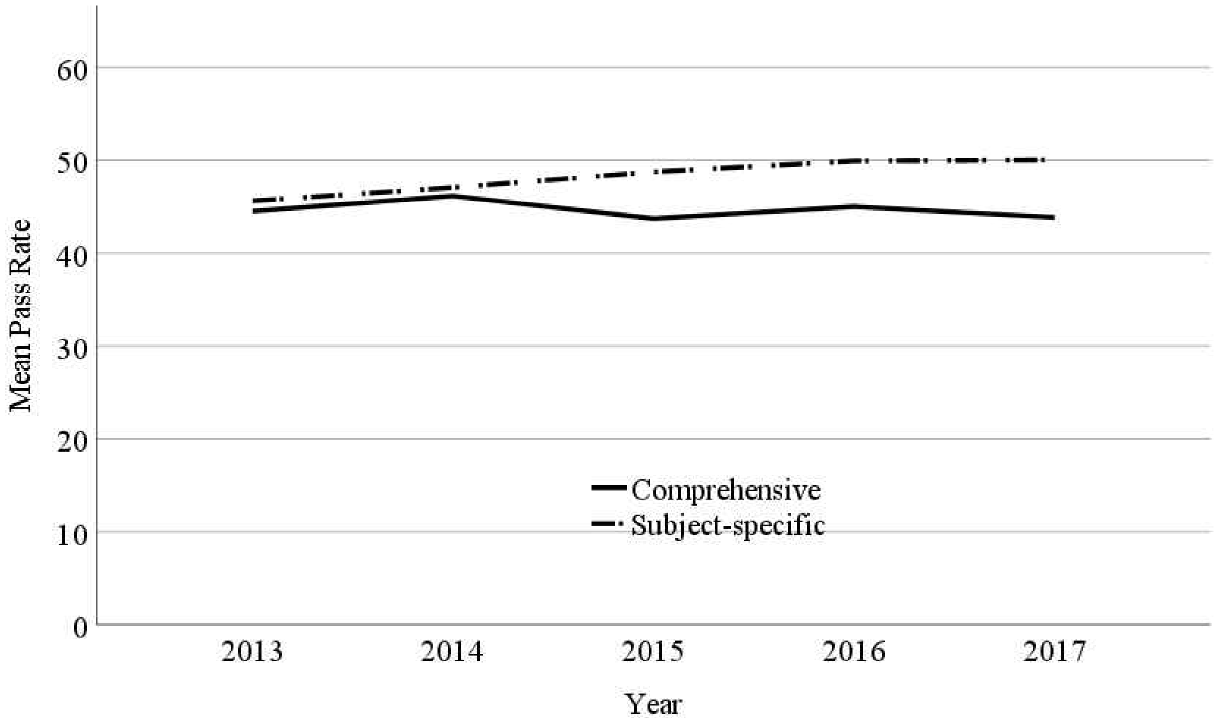

Figure 18. Changes in school district mean pass rates, by school district group.

The boxplot in Figure 19 shows the dispersions of pass rates were greater, and increased each year, for the comprehensive group. Most of the low outliers were in the comprehensive group. The only high outliers were in the subject-specific group. The low and high outliers were also outliers in school district mean scale scores for all students.

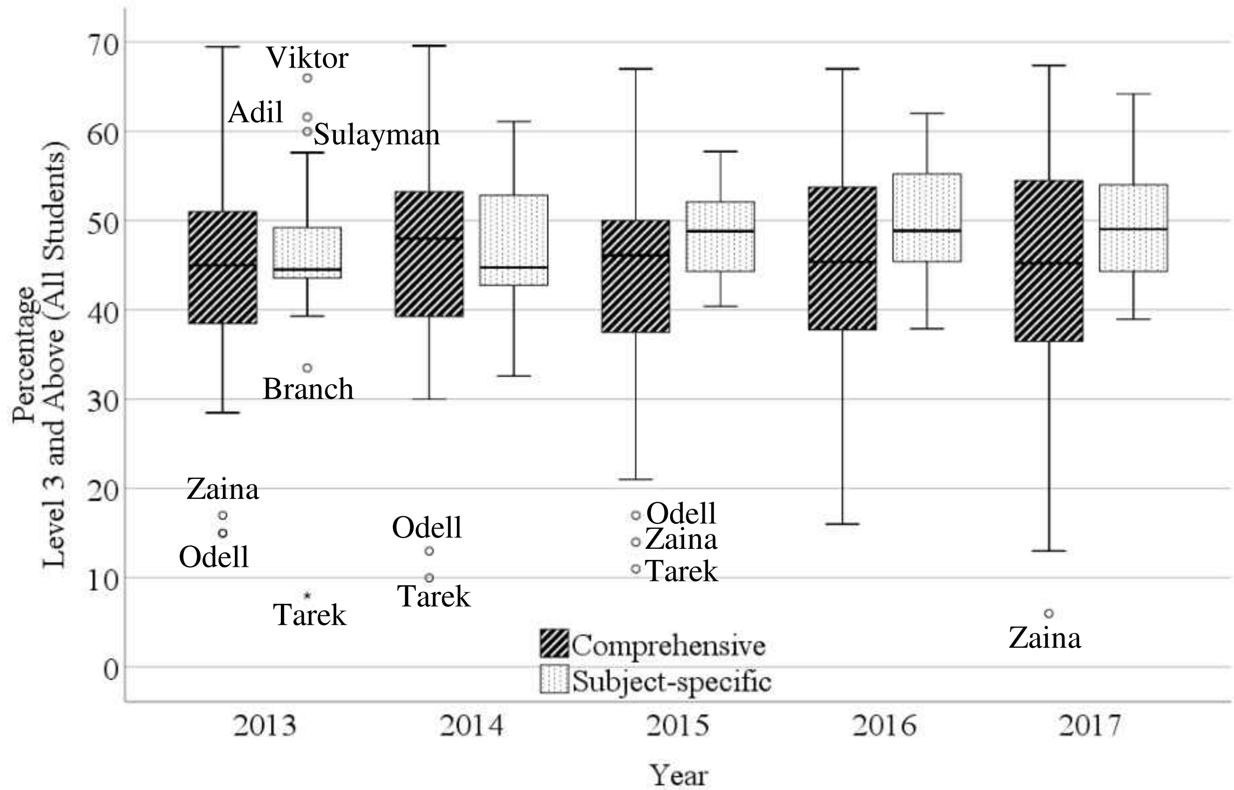

Figure 19. Boxplot of 2013–2017 school district pass rates for all students, by school district group. Minimum passing score is 203, which equates to Achievement Level 3 (FLDOE, 2017o).

To test for statistical significance in the differences in mean pass rates for all students between the two school district groups, independent samples *t*-tests were conducted. These tests were conducted only for the school district mean pass rates for all students. Pass rates for low SES students and ELs were not included in the data collection procedures.

All samples met the assumptions of normality of distribution. The samples for 2015 and 2017 did not meet the assumption for equality of variance, so the degrees of freedom were adjusted. As shown in Table 79, the pass rate for the subject-specific group of school districts was significantly higher than the comprehensive group in 2015 (d = -0.54) and 2017 (d = -0.57). These were medium effect sizes according to Cohen’s (1988) guidelines.

Table 79

2013–2017 Independent Samples Tests of School District Mean Pass Rates

		Descriptive statistics							Paired samples *t*-tests				
												95% CI	
Year	School district group	*n*	*M*	*Mdn*	*SD*	*SEM*	Skew	Kurt	*t*	*df*	*p*	*LL*	*UL*
2013	C	49	44.52	45.00	11.93	1.70	-0.53	0.63	-0.32	65	.74	-7.77	5.55
	S	18	45.63	44.52	12.60	2.97	-1.27	4.13					
2014	C	52	46.16	48.00	11.29	1.57	-0.94	1.93	-0.29	65	.78	-7.24	5.43
	S	15	47.06	44.75	8.96	2.31	0.27	-0.83					
2015	C	53	43.72	46.10	11.98	1.65	-0.75	0.76	-2.29	47.80	.03	-9.45	-0.61
	S	14	48.76	48.81	5.46	1.46	0.12	-0.86					
2016	C	52	45.02	45.38	11.82	1.64	-0.43	-0.11	-1.52	65	.13	-9.95	0.13
	S	15	49.92	48.87	7.26	1.87	0.08	-0.92					
2017	C	49	43.86	45.25	13.73	1.96	-0.63	0.20	-2.43	57.02	.02	-11.25	-1.09
	S	17	50.03	49.05	6.63	1.61	0.42	-0.21					

Because of the different sample sizes and varying demographics of the independent samples, paired samples *t*-tests were conducted to test the differences in school district mean pass rates for all students between paired samples of school districts from both groups. School districts were matched by school district student population, low SES student percentage, and percentage of ELs. These were the same paired samples of school districts used for Research Question 2.

All samples met the assumptions of normality of distribution. The assumption of homogeneity of variance was satisfied by Pitman-Morgan tests for each paired sample (Gardner, 2001; Morgan, 1939; Pitman, 1939). As shown in Table 80, the paired samples *t*-tests showed no statistically significant difference in mean pass rates between the two groups of school districts.

Table 80

2013–2017 Paired Samples t-Tests of School District Mean Pass Rates

		Descriptive statistics							Paired samples *t*-tests				
												95% CI	
Year	School district group	*n*	*M*	*Mdn*	*SD*	*SEM*	Skew	Kurt	*t*	*df*	*p*	*LL*	*UL*
2013	C	18	48.05	50.00	10.07	2.37	-1.43	5.23	1.14	17	.27	-2.08	6.92
	S	18	45.63	44.52	12.60	2.97	-1.27	4.13					
2014	C	15	50.79	49.80	8.66	2.24	-0.05	1.27	1.27	14	.26	-2.57	10.02
	S	15	47.06	44.75	8.96	2.31	0.27	-0.83					
2015	C	14	47.84	48.77	10.32	2.76	-2.21	6.26	-0.29	13	.78	-7.71	5.88
	S	14	48.76	48.81	5.46	1.46	0.12	-0.87					
2016	C	15	51.01	51.40	6.86	1.77	-0.66	0.45	0.42	14	.68	-4.44	6.61
	S	15	49.92	48.87	7.26	1.87	0.08	-0.92					
2017	C	17	48.93	49.25	11.91	2.89	-0.72	0.42	-0.29	16	.77	-9.05	6.85
	S	17	50.03	49.05	6.63	1.61	0.42	-0.21					

To test if the 2013 to 2017 changes in mean pass rates for all students were statistically significant for either school district group, independent samples *t*-tests were conducted. These tests compared each school district's 2013 mean pass rate to its 2017 mean pass rate. As shown in Table 81, the comprehensive group mean pass rate for all students decreased numerically from 2013 to 2017. Both the 2013 and 2017 distributions were sufficiently normal for the purposes of conducting a *t*-test, i.e., skew < |2.0| and kurtosis < |9.0|; (Schmider et al., 2010). Levene's test, $F(96) = 1.01$, $p = .31$, confirmed the equality of variance. The independent samples *t*-test showed no statistically significant change in the comprehensive group mean pass rate for all students from 2013 to 2017.

Table 81

2013 vs. 2017 School District Mean Pass Rate for All Students, Comprehensive Group

Descriptive statistics								Independent samples test				
											95% CI	
Year	*n*	*M*	*Mdn*	*SD*	*SEM*	Skew	Kurtosis	*t*	*df*	*p*	*LL*	*UL*
2013	49	44.52	45.00	11.93	1.70	-0.53	0.63	0.26	96	.80	-4.49	5.83
2017	49	43.85	45.25	13.73	1.96	-0.63	0.20					

The school district mean pass for the subject-specific group increased numerically from 2013 to 2017, as shown in Table 82. Both distributions were sufficiently normal for the purposes of conducting a *t*-test, i.e., skew < |2.0| and kurtosis < |9.0|; (Schmider et al., 2010). Levene's test, $F(33) = 1.13$, $p = .30$, confirmed equality of variance. The independent samples *t*-test, also shown in Table 82, showed the 4.4 percentage point increase was not statistically significant.

Table 82

2013 vs. 2017 School District Pass Rate Comparison, Subject-specific Group

Descriptive statistics								Independent samples test				
											95% CI	
Year	*n*	*M*	*Mdn*	*SD*	*SEM*	Skew	Kurtosis	*t*	*df*	*p*	*LL*	*UL*
2013	18	45.63	44.52	12.60	2.97	-1.27	4.13	-1.28	33	.21	-11.38	2.59
2017	17	50.03	49.05	6.63	1.61	0.42	-0.21					

Analyses of Mean Raw Scores by Subject Area

To what extent did the 2013–2017 8th grade FCAT 2.0 Science/SSA school district mean raw scores for all students, by subject area (nature of science, physical science, life science, and Earth/space science), differ between the two groups of school districts?

Independent samples *t*-tests were conducted to examine differences in 2013 through 2017 8th grade FCAT 2.0 Science/SSA school district mean raw scores for all students, by subject area, between the two groups of school districts. (Mean raw scores for low SES students and ELs were not included in the data). The purpose was to search for differences between the two school district groups that may not be apparent in the mean scale score analyses. The raw score represents the school district mean number of assessment items students answered correctly in each subject area (nature of science, physical science, life science, and Earth/space science). Scale scores are derived from the raw scores, with items of higher complexity counting for more scale points than lower complexity items. The *t*-tests determined if there was a statistically significant difference raw scores for any of the four subject areas, between the two groups of school districts. The *t*-tests were non-directional, $\alpha = .05$. Positive *t*-values indicate the mean raw scores of the comprehensive group were higher than those of the subject-specific group, and vice-versa.

As shown in Table 83, the raw score distributions for each group were sufficiently normal for the purposes of conducting a *t*-test, i.e., skew $< |2.0|$ and kurtosis $< |9.0|$ (Schmider et al., 2010). As shown in Table 84, the assumption of equality of variances was not satisfied by the Levene's tests for any of the four subject areas, so the degrees of freedom were adjusted accordingly (George & Mallery, 2010). The boxplot in Figure 20 shows the dispersion of raw scores for each subject area was greater for the comprehensive group than for the subject-specific group. Low outliers were evident in both groups. One low outlier, Tarek school district, was a low outlier in every subject area before and after changing from subject-specific to comprehensive science from 2013 to 2014.

Table 83

Distributions of 2013–2017 School District Mean Raw Scores, by Subject Area

School district group	Subject area	Maximum raw score points possible	*N*	*M*	*Mdn*	*SD*	*SEM*	Skew	Kurtosis
C	Nature of science	11	255	6.69	6.79	0.71	0.04	-0.51	0.57
	Physical science	15	255	9.31	9.40	0.96	0.06	-0.71	0.86
	Life science	15	255	8.97	9.00	0.90	0.06	-0.59	0.86
	Earth/space science	15	255	9.18	9.20	0.96	0.06	-0.67	0.96
S	Nature of science	11	79	6.85	6.90	0.58	0.07	-1.37	6.15
	Physical science	15	79	9.61	9.73	0.73	0.08	-1.52	6.50
	Life science	15	79	9.30	9.25	0.58	0.07	-0.56	2.49
	Earth/space science	15	79	9.19	9.16	0.65	0.07	-0.28	1.03

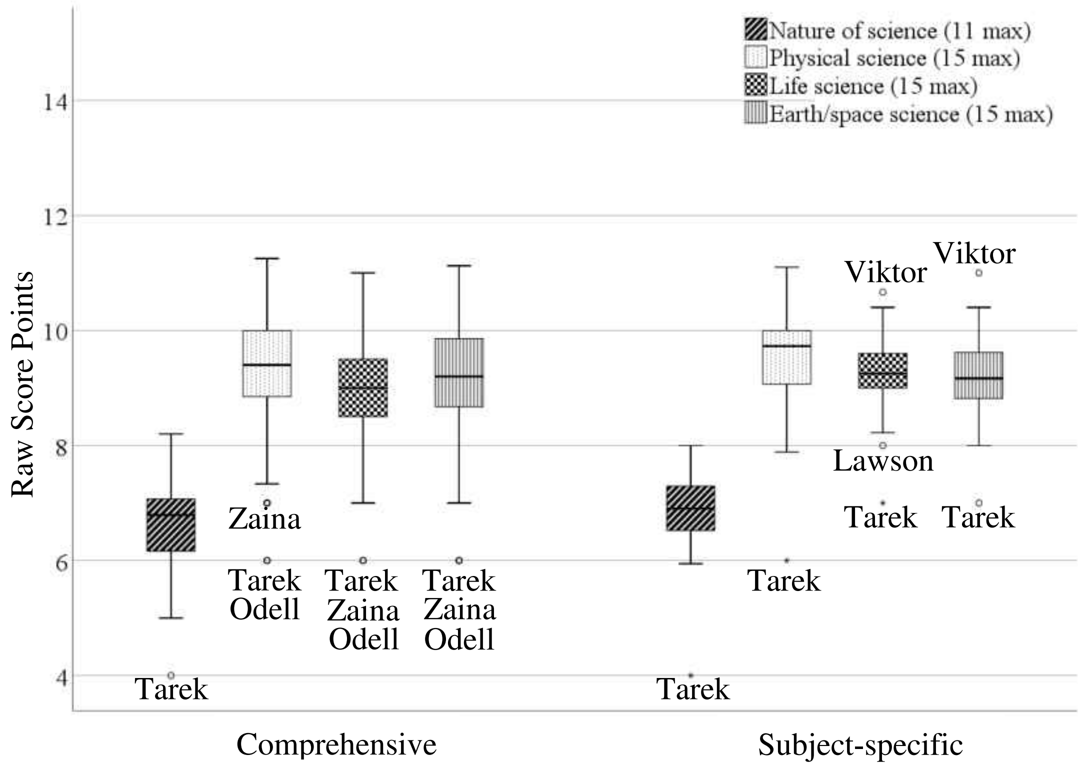

Figure 20. Boxplot of 2013–2017 school district mean raw scores, by subject area. Nature of science raw scores appear lower due to fewer number of points possible (11).

As shown in Table 84, the independent samples *t*-test revealed that the subject-specific group had statistically significantly higher mean raw scores for nature of science, physical science, and life science, than the comprehensive group. The effect sizes for these differences, also shown in Table 84, were small-to-medium, based on Cohen's (1988) guidelines. There was no statistically significant difference in the school district mean raw score for Earth/space science between the two groups of school districts. The boxplot revealed low outliers for both groups of school districts in every subject area, and two high outliers for the subject-specific group in life science and Earth/space science subject areas.

Table 84

Independent Samples Tests of School District Raw Scores by Subject Area

	Levene's test		Independent samples *t*-test					
						95% CI		
Subject area	F	*p*	*t*	*df*	*p*	LL	UL	*d*
Nature of science	5.43	.02	-2.07	157.21	.04	-0.32	-0.01	-0.24
Physical science	7.41	.007	-2.95	167.80	.004	-0.50	-0.10	-0.33
Life science	10.67	.001	-3.87	203.46	<.001	-0.50	-0.16	-0.39
Earth/space science	9.49	.002	-0.08	191.83	.94	-0.19	0.18	NA

Due to the outliers revealed in the boxplot, Mann-Whitney tests were conducted on the school district mean raw score data. As shown in Table 85, according to the Mann-Whitney tests, only physical science and life science mean raw scores were significantly different between the two groups of school districts. The subject-specific group had a higher 2013–2017 school district mean raw score for physical science (small to medium effect size) and life science (medium effect size) than the comprehensive group, based on Cohen’s (1988) guidelines.

Table 85

Mann-Whitney Tests of School District Mean Raw Scores, by Subject Area

	Subject area			
Statistic	Nature of science	Physical science	Life science	Earth/space science
U	8734.00	8204.50	7528.00	9774.50
Z	-1.79	-2.50	-3.40	-0.40
p	.073	.013	.001	.690
d	NA	-0.29	-0.41	NA

Correlation of School District Student Population to Mean Scale Scores

To what extent was school district student population correlated with the 2013–2017 8th grade FCAT 2.0 Science/SSA school district mean scale scores for all students, low SES students, regardless of the science course type offered? Pearson *r* correlation tests were conducted to determine the correlation of school district student population to the 2013 through 2017 8th grade FCAT 2.0 Science/SSA school district mean scale scores for all students, low SES students, and ELs, without regard to the type of science courses offered by the school districts.

As shown in Table 86 and the scatterplot in Figure 21, the correlation of school district student population to the 2013–2017 8th grade FCAT 2.0 Science/SSA school district mean scale score for all students was not statistically significant (Steinberg, 2011). Also shown in Table 86, and the scatterplot in Figure 22, the correlation of school district student population to the 2013 through 2017 8th grade FCAT 2.0 Science/SSA school district mean scale score for low SES students was statistically significant, and medium in strength (Steinberg, 2011). School districts with larger student populations had lower school district mean scale scores for low SES students than school districts with smaller student populations. Finally, as shown in Table 86 and the scatterplot in Figure 23, the correlation of school district student population to the 2013 through 2017 8th grade FCAT 2.0 Science/SSA school district mean scale score for ELs was not statistically significant (Steinberg, 2011).

Table 86

Correlation of School District Student Population to Mean Scale Scores for Student Groups

Student group	*r*	*df*	*p*
All students	-.02	334	.74
Low SES students	-.33	334	<.001
ELs	.09	179	.21

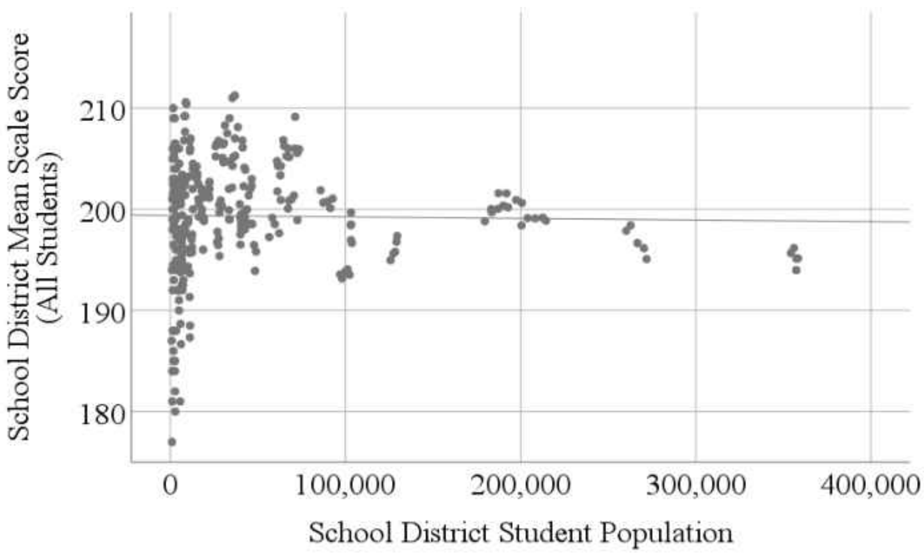

Figure 21. Correlation of 2013–2017 school district mean scale score for all students to school district student population. r = -.02, N = 334, p = .74.

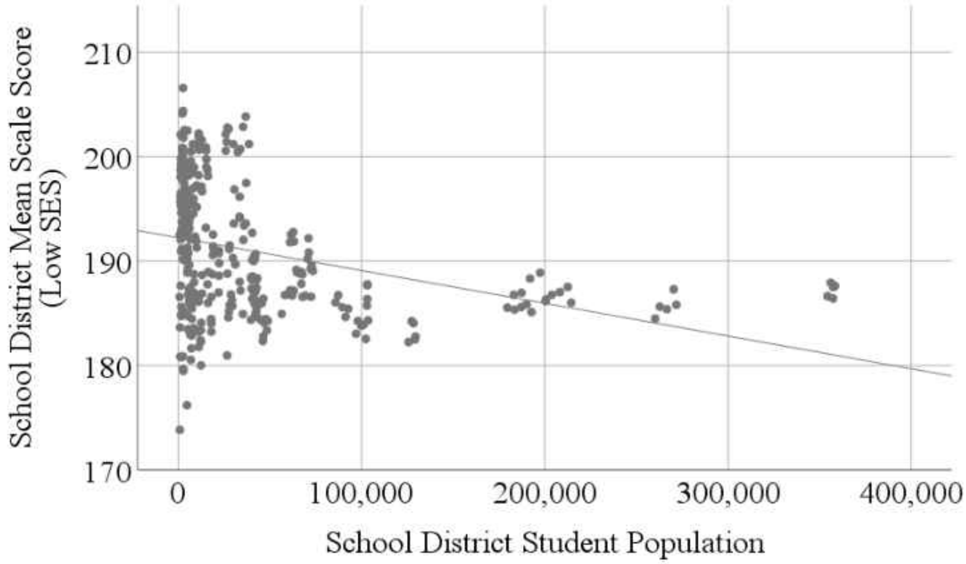

Figure 22. Correlation of 2013–2017 school district mean scale score for low SES students to school district student population. $r = -.33$, $N = 334$, $p < .001$.

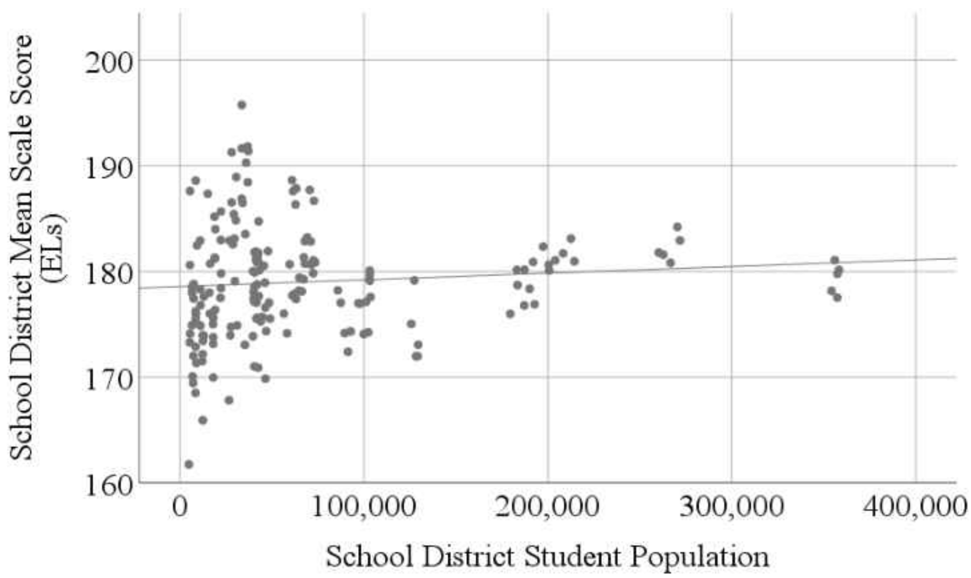

Figure 23. Correlation of 2013–2017 school district mean scale score for ELs to school district student population. $r = .09$, $N = 334$, $p = .21$.

Analyses of Mean Scale Scores of School Districts that Changed Course Offering

Did changing the type of course offering have any general impact on school district mean scale scores for all students? This question was answered qualitatively by graphing the 2013–2017 school district mean scale scores of the seven school districts that changed course offerings during these five years, as shown in Figure 24. The descriptive statistics for these seven school districts are shown in Table 87. Of the three school districts that changed from comprehensive to subject-specific courses (Lucas, Emmett, and Mirela), the mean scale scores increased in Emmett and Mirela, and declined in Lucas in the year following the change. In Emmett school district, the mean scale score declined in the second year following the change. Of the four school districts that changed from subject-specific to comprehensive courses (Gottfried, Earline, Tarek, and Branch), mean scale scores increased in Gottfried, Tarek, and Branch, and declined in Earline in the year following the change. Only in Gottfried did the increase remain through years two to four following the change. The smallest school district, Tarek, had the most fluctuation in school district mean scale scores in the years following the change. In the largest school district, Branch, the mean scale scores remained relatively stable in the years following the change.

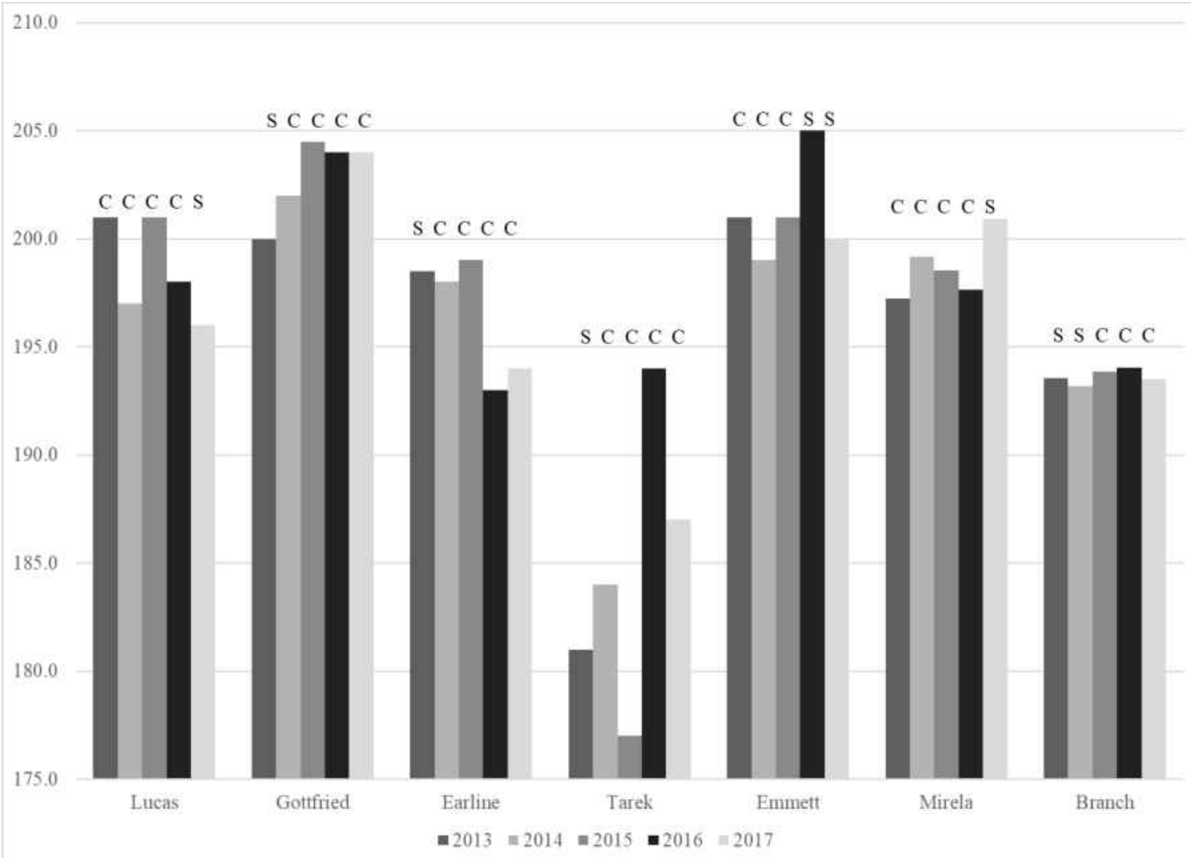

Figure 24. Mean scale scores for all students of school districts that changed science course offerings, 2013–2017. C = comprehensive, S = subject-specific.

Table 87

Descriptive Statistics of School Districts that Changed Science Course Offerings, 2013–2017

School district	Student population	Low SES percentage	EL percentage
Branch	99,656	59.23	10.79
Earline	1,910	57.88	0.64
Emmett	12,380	60.32	7.12
Gottfried	12,855	58.17	2.54
Lucas	4,991	54.55	0.26
Mirela	59,858	65.87	18.42
Tarek	889	72.20	2.28
M	25,913.83	61.17	6.01

Summary

The problem underpinning this study was the stagnation of science proficiency of 8th grade science students as measured by the Florida Comprehensive Assessment Test 2.0 for Science (FCAT 2.0 Science)/Statewide Science Assessment (SSA). The purpose of the tests was to determine if there were statistically significant differences in student achievement on Florida's 8th grade FCAT 2.0 Science/SSA for three groups of students between two groups of school districts: those that offered comprehensive middle grades science courses and those that offered subject-specific middle grades science courses). The three student groups were all students, low SES students, and ELs.

Research Question 1 used all available 2013, 2014, 2015, 2016 and 2017 8th grade FCAT 2.0 Science/SSA mean scale score data, with no controls for demographic factors that may have affected the mean scale scores. Levene's tests were used to test the assumption of equality of variance in the samples. In the independent samples of 2013 school district mean scale scores for all students and low SES students, and the 2015 mean scale scores for ELs, the variance was statistically significantly greater for the comprehensive group than for the subject-specific group. The dispersions of all samples were greater for the comprehensive group.

Numerically, the school district mean scale scores for all students and for ELs were higher for the subject-specific group. The school district mean scale scores for low SES students were numerically higher for the comprehensive group. Independent samples *t*-tests found no statistically significant differences in the 2013, 2014, 2015, 2016, or 2017 school district mean scale scores for any of the three student groups, between the two groups of school districts. The tests of the 2015, 2016, and 2017 school district mean scale scores for all students approached significance ($p < .10$), with the subject-specific group mean scale scores slightly higher than the comprehensive group. The test of the 2014 school district mean scale score for low SES students approached significance, with the comprehensive group mean scale score slightly higher than the subject specific group.

The tests conducted for Research Questions 2, 3, and 4 used paired samples of school districts from each of the two groups, matched by student population, percentage of low SES students, and percentage of ELs. The dispersions of the school district mean scale scores for all students were greater for the comprehensive group than for the subject-specific group. In four of the five years, the comprehensive group mean scale scores for all students were numerically higher than the subject-specific group. The tests conducted for Research Question 2 found no statistically significant differences in the 2013, 2014, 2015, 2016, or 2017 school district mean scale scores for all students between the two groups of school districts in the paired samples. None of the Research Question 2 tests approached significance.

The dispersion of the school district mean scale scores for low SES students in the paired samples was greater each year for the comprehensive group. The school district mean scale scores for low SES students were numerically greater for the comprehensive group in four of the five years analyzed. Paired samples t-tests conducted for Research Question 3 found no statistically significant differences in the 2013, 2014, 2015, 2016, or 2017 school district mean scale scores for low SES students between the two groups of school districts in the paired samples. The test of the 2014 school district mean scale scores for low SES students approached significance, with a p-value just over the $\alpha = .05$ threshold, and the comprehensive group mean scale score for low SES students higher than that of the subject-specific group.

The dispersion of the school district mean scale scores for ELs in the paired samples was greater each year for the comprehensive group. The school district mean scale scores for ELs were numerically greater for the comprehensive group in four of the five years analyzed. The tests for Research Question 4 found no statistically significant differences in the 2013, 2014, 2015, 2016, or 2017 school district mean scale scores for ELs between the two groups of school districts the paired samples. The test of the 2016 school district mean scale scores for ELs approached significance. The *p*-value was just over the $\alpha = .05$ threshold, with the comprehensive group mean scale score for low SES students higher than that of the subject-specific group.

Seven additional analyses were conducted to check the validity and interpret the results of the Research Questions 1–4 tests. First, school-level mean scale scores for all students were analyzed to check the validity of school district mean scale score tests. The dispersion of the school-level mean scale scores was greater each year for the comprehensive group. The school-level mean scale scores were numerically greater each year for the subject-specific group. Independent samples *t*-tests of school-level mean scale scores for all students supported the Research Question 1 tests. There were no statistically significant differences found in the 2013, 2014, 2015, 2016, or 2017 8th grade FCAT 2.0 Science/SSA school-level mean scale scores for all students, between the two school district groups.

Second, due to unequal variances in some independent and paired samples, additional non-parametric tests were conducted on the samples used for Research Questions 1–4. Mann-Whitney tests of the independent school district samples corroborated the results of the Research Question 1 independent samples *t*-tests. No statistically significant differences were found in the

2013, 2014, 2015, 2016, or 2017 FCAT 2.0 Science/SSA school district mean scale scores for the three groups of students (all students, low SES students, or ELs), between the two groups of school districts.

Wilcoxon signed rank tests of the paired school district samples corroborated the paired samples *t*-tests conducted for Research Questions 2, 3, and 4. No statistically significant differences were found in the 2013, 2014, 2015, 2016, or 2017 FCAT 2.0 Science/SSA school district mean scale scores for the three groups of students (all students, low SES students, or ELs) between the two groups of school districts.

Third, independent samples *t*-tests analyzed changes in school district mean scale scores from 2013 to 2017. These tests found a statistically significant increase from 2013 to 2017 in the school district mean scale score for ELs in the comprehensive group of school districts. Further analysis of this increase revealed that a single low outlier in 2013 and a single high outlier in 2017. When the *t*-tests were conducted with the outliers removed from the samples, the change was not statistically significant. No other significant 2013 to 2017 changes were found in the 8th grade FCAT 2.0 Science/SSA school district mean scale scores for three groups of students in either group of school districts.

Fourth, independent samples *t*-tests were used to look for statistically significant differences in the school district mean pass rates for all students between the two school district groups. In 2015 and 2017, the subject-specific group mean pass rates were statistically significantly higher than those of the comprehensive group. Numerically, from 2013 to 2017, the school district mean pass rate for the subject-specific group increased, while that of the comprehensive group decreased. However, independent samples *t*-tests found these 2013 to 2017 changes in pass rates were not statistically significant for either group of school districts.

Fifth, additional analyses were conducted to determine if there were statistically significant differences in the overall 2013–2017 school district mean raw scores for all students, by subject area, between the two groups of school districts. Independent samples *t*-tests showed that the subject-specific group had significantly higher mean raw scores than the comprehensive group in the nature of science, physical science, and life science categories. The effect sizes of these differences were small to medium. No statistically significant difference was found for the Earth/space science subject area between the two groups of school districts. Levene's tests showed the variance of mean raw scores was statistically significantly greater for the comprehensive group in all four categories. Because of these differences in variance, the *t*-test results were checked for validity with Mann-Whitney *U* tests. These tests corroborated the *t*-test results for two subject areas. The subject-specific group had statistically significantly higher school district mean raw scores than the comprehensive group in physical science and life science. The effect sizes of these differences were small to medium. No significant differences were found in the school district mean raw scores for nature of science or Earth/space science between the two groups of school districts.

Sixth, additional tests were conducted to determine if school district student population was statistically significantly correlated with school district mean scale scores for the three groups of students, regardless of the type of science course offered. Pearson *r* correlation tests found no statistically significant correlation between school district student population and the 2013–2017 8^{th} grade FCAT 2.0 Science/SSA school district mean scale scores for all students or for ELs, regardless of the science course type offered. A statistically significant, medium, negative correlation was found between school district student population and the 2013–2017 8th grade FCAT 2.0 Science/SSA school district mean scale scores for low SES students, regardless of the science course type offered. From 2013 to 2017, school districts with larger student population had moderately lower school district mean scale scores for low SES students than smaller school districts, regardless of the type of science course offered.

Finally, the mean scale scores of school districts that changed course offerings during the years 2013–2017 were examined qualitatively. There were no consistent increases or decreases in school district mean scale scores following a change in the type of science course offering. Smaller school districts appeared to have had wider fluctuations in mean scale scores following a change in course offering, while the mean scale scores of larger school districts remained relatively stable in the years following a change.

This chapter has presented the findings of the quantitative analyses performed in this study. Tests were conducted to answer four research questions and six additional questions. The next chapter presents a discussion of these findings, implications for practice, recommendations for further research, and conclusions of this study.

CHAPTER V: CONCLUSION

Introduction

In the preceding chapter, the results of the quantitative analyses of data were reported. The research questions and additional questions all focused on different methods of examining the differences in science achievement of three groups of students, between the two school district groups.

This chapter begins with a summary of the background of the research, including the problem statement, purpose of the research, and the methods used to conduct the research. The chapter then moves to a summary of the findings of the study. Following that is a discussion of the findings, organized by the themes presented in the conceptual framework. Next, the implications of the study as they relate to educational leadership are presented. The chapter concludes with recommendations for further research related to comprehensive and subject-specific science education.

Summary of the Study

The problem that generated this study was the stagnation of science proficiency of 8th grade science students as measured by required assessments in Florida. From 2013 to 2017, less than half of all 8th grade students achieved the minimum level of proficiency on the FCAT 2.0 Science/ SSA (Florida Department of Education [FLDOE], 2012b, 2015e, 2016c). The results of similar national and international science assessments mirrored this stagnation of student science achievement (FLDOE, 2017k; International Association for the Evaluation of Educational Achievement [IEA], 2016; Kastberg et al., 2016; National Center for Education Statistics

[NCES], 2017b). To improve science achievement, many Florida school districts have followed a national trend in moving away from traditional, subject-specific science courses to integrated, or comprehensive science courses (Banilower et al., 2013; Czerniak, 2007; FLDOE, 2017d; Hoeg & Bencze, 2017; Huff & Yager, 2016; National Research Council [NRC], 2012). Comprehensive courses are recommended to improve middle grades science achievement because they focus on concepts that span the traditional subject area boundaries, helping students learn the core ideas of each subject area in more depth (NRC, 2012; Rutherford, 1990). However, scant quantitative research exists to show if either comprehensive or subject-specific science courses improve student achievement (Åström & Karlsson, 2012; Tamassia & Frans, 2014).

The purpose of this study was to determine if there was a difference in student achievement on the 8th grade FCAT 2.0 Science/SSA from 2013 to 2017, between two groups of school districts: those that offered comprehensive middle grades science courses (comprehensive group) and those that offered subject-specific middle grades science courses (subject-specific group). The school district mean scale scores for three groups of students were compared using multiple levels of quantitative analyses. The three student groups were (a) all assessed 8th grade students (all students), (b) low socioeconomic status (SES) students, and (c) English learners (ELs). Additional analyses were conducted comparing school-level mean scale scores, school district mean pass rates, and school district mean raw scores for each science subject area.

Five concepts comprised the framework of this study and guided the literature review. First, the status of science education in the US and the state of Florida was reviewed to support the need for this study. Results of international, national, and state standardized assessments, including Program for International Student Assessment (PISA), Trends in International Mathematics and Science Study (TIMSS), National Assessment of Educational Progress (NAEP), and the FCAT 2.0 Science)/SSA showed that science achievement of middle school students in the US has been stagnant since 2013 (FLDOE, 2017k; IEA, 2016; Kastberg et al., 2016; NCES, 2017b). The implications of these scores were explored from national, student, and Florida educational leadership perspectives.

Standardized assessment of student achievement on middle grades science standards was the second concept supporting this study. The frameworks and science standards assessed in the PISA, TIMSS, NAEP, and FCAT 2.0 Science/SSA science assessments were compared. The implications of standardized assessments on student performance were explored.

The third concept underpinning this study was related to science standards at the international, national, and state levels. The history of the development of science standards was explored. Documentation on the consistency and alignment of Florida's NGSSS for science to FCAT 2.0 Science/SSA was reviewed.

The challenges school districts encounter when implementing new science curriculum and course offerings comprised the fourth concept underpinning this study. These challenges were the basis of the controls used in the Research Questions 2, 3, and 4 tests, and the reason for analyzing separately the scores for three groups of students. Larger school districts often experience inconsistency in the implementation of new course offerings due to varying attitudes, beliefs, education, professional development, and subject area certifications among the teaching force (Davis, 2003; Diehl, 2005). Student factors such as poverty and lack of English language proficiency have been shown to influence student performance on standardized assessments, presenting special challenges for school districts with high percentages of low SES students and ELs (Amaral et al., 2002; Driscoll et al., 2003; Lippman et al., 1996; Maerten-Rivera et al., 2016; Matkins et al., 2014; Wiseman, 2012).

The fifth concept, and the focus of this study, was science courses that support student learning. Extensive searches yielded only two published, peer-reviewed, academic studies, four dissertations, and one master's thesis that compared comprehensive to subject-specific science courses (Åström & Karlsson, 2012; Tamassia & Frans, 2014). These relevant empirical studies were reviewed.

This study included four research questions focused on determining if there was a significant difference in student science achievement between the two groups of school districts. Seven additional questions arose during the research question analyses, and the data were analyzed to answer these questions. The research questions are listed below, followed by the additional questions.

Research Questions

1. To what extent did 8^{th} grade FCAT 2.0 Science/SSA school district mean scale scores for 2013, 2014, 2015, 2016, and 2017 differ between two groups of Florida school districts: those that offered comprehensive science courses and those that offered subject-specific science courses?

2. To what extent did 8^{th} grade FCAT 2.0 Science/SSA school district mean scale scores for 2013, 2014, 2015, 2016, and 2017 differ between two groups of Florida school districts: those that offered comprehensive science courses and those that offered subject-specific science

courses, when the school districts are matched by overall population size, low SES student percentage, and EL percentage?

3. To what extent did 8th grade FCAT 2.0 Science/SSA school district mean scale scores for low SES students for 2013, 2014, 2015, 2016, and 2017 differ between two groups of Florida school districts: those that offered comprehensive science courses and those that offered subject-specific science courses, when the school districts are matched by overall population size, low SES student percentage, and EL percentage?

4. To what extent did 8th grade FCAT 2.0 Science/SSA school district mean scale scores for ELs for 2013, 2014, 2015, 2016, and 2017 differ between two groups of Florida school districts: those that offered comprehensive science courses and those that offered subject-specific science courses, when the school districts are matched by overall population size, low SES student percentage, and EL percentage?

Additional Questions

1. To what extent did 8th grade FCAT 2.0 Science/SSA school-level (as opposed to school district-level) mean scale scores for all students for 2013, 2014, 2015, 2016, and 2017, differ between two groups of Florida schools: those that offered comprehensive science courses and those that offered subject-specific science courses?

2. Do non-parametric tests of the Research Questions 1, 2, 3, and 4 samples indicate any statistically significant differences in the 2013, 2014, 2015, 2016, or 2017 8th grade FCAT 2.0 Science/SSA school district mean scale scores for the three groups of students between the two groups of school districts?

3. To what extent did the 8th grade FCAT 2.0 Science/SSA school district mean scale scores for the three groups of students change, from 2013 to 2017, for each group of school districts?

4. To what extent did the 2013, 2014, 2015, 2016, or 2017 8th grade FCAT 2.0 Science/SSA school district pass rates (percentage of students who achieved Level 3 or higher) for all students differ between the two groups of school districts?

5. To what extent did the 2013–2017 8th grade FCAT 2.0 Science/SSA school district mean raw scores for all students, by subject area (nature of science, physical science, life science, and Earth/space science), differ between the two groups of school districts?

6. To what extent was school district student population correlated with the 2013–2017 8th grade FCAT 2.0 Science/SSA school district mean scale scores for all students, low SES students, regardless of the of science course type offered?

7. Did changing the type of course offering have any general impact on school district mean scale scores for all students?

The questions were answered both quantitatively and, for the final additional question, qualitatively. The quantitative analyses began with parametric tests offering the greatest degrees of freedom and the fewest controls and progressed through tests that traded degrees of freedom for increased controls. Alternate non-parametric tests were performed to confirm the validity of the results due to outliers and differences in dispersion present in the data. Further quantitative analyses were conducted to search for differences that may have been concealed in the tests of school district mean scale scores. Overall qualitative observations of the data were made to interpret the findings and search for differences between the two groups not evident in the quantitative analyses.

Summary of Findings

The major findings of this study were categorized by differences in mean scale scores, mean pass rates, mean raw scores, and demographics. These findings were:

Mean Scale Scores

1. For both school district groups, mean scale scores for all student groups were lower than the minimum passing score from 2013 to 2017.

2. The comprehensive group had greater dispersions of mean scale scores for all student groups in all independent and paired school district and school-level samples. The greater dispersions were bi-directional. Each year, the comprehensive group had both the maximum and minimum school district mean scale scores. Many of these were found to be statistically significant differences in variance.

3. In independent samples of all school districts in both groups, the subject-specific group mean scale scores for all students and for ELs were numerically higher. This difference approached statistical significance in 2015, 2016, and 2017. Conversely, the mean scale scores for low SES students were numerically higher for the comprehensive group in the independent samples. None of these differences approached significance.

4. In paired samples of school districts matched by student population, percentage of low SES students, and percentage of ELs, the mean scale scores for all three student groups were

numerically higher for the comprehensive group. None of these differences approached significance.

5. For the seven school districts that changed course offerings from 2013 to 2017, there were no consistent increases or decreases in mean scale scores following the change. In the years following the change, smaller school districts tended to have greater fluctuations in mean scale scores than larger school districts.

6. Overall, no statistically significant differences were found in the school district mean scale scores between the two school district groups, for any of the three student groups. Several tests approached significance in independent samples, but in paired samples of school districts matched by student population, percentage of low SES students, and percentage of ELs, none of the test results approached significance.

Mean Pass Rates

1. In independent samples *t*-tests, the 2015 and 2017 school district mean pass rates for all students were statistically significantly higher for the subject-specific group than for the comprehensive group. However, in paired samples *t*-tests, there were no statistically significant differences in the 2013, 2014, 2015, 2016, or 2017 school district mean pass rates for all students between the two school district groups.

2. For the first time since before 2013 for either school district group, the 2017 subject-specific group school district mean pass rate for all students exceeded fifty percent.

3. From 2013 to 2017, the school district mean pass rate for all students for the subject-specific group increased numerically, while that for the comprehensive group decreased numerically. By 2017, the subject-specific group mean pass rate was 6.17 percentage points

higher than that of the comprehensive group. Independent samples *t*-tests showed the 2013 to 2017 change in pass rates was not statistically significant for either school district group.

4. The comprehensive group had greater dispersions of mean pass rates for all students than the subject-specific group. Many of these were statistically significant differences in variance. The greater dispersions were bi-directional. Both the maximum and minimum school district mean pass rates for the comprehensive group exceeded those of the subject-specific group. Additionally, the dispersion of the school district mean pass rates for all students for the comprehensive group increased from 2013 to 2017, while that for the subject-specific group decreased.

Mean Raw Scores

1. School district mean raw scores for all students were statistically significantly higher for the subject-specific group than for the comprehensive group in the subject areas of nature of science, life science, and physical science.

2. There was no statistically significant difference between the two school district groups in mean raw scores for the Earth/space science subject area.

Demographics

1. For both school district groups, the achievement gap between all students and low SES students was statistically significant, with a very large effect size (Sawilowsky, 2003).

2. For both school district groups, the achievement gap between all students and ELs was statistically significant, with a huge effect size (Sawilowsky, 2003).

3. School district student population had a statistically significant, medium, negative correlation with the school district mean scale scores for low SES students, regardless of science course offering. Larger school districts tended to have lower mean scale scores for low SES students. Of the 19 consistently low science-achievement school districts (2013–2017 mean scale scores below 197), all had higher than average student populations, percentages of low SES students, and percentages of ELs. These were comprised of 4 of the state's 10 largest school districts (including the two largest); 9 of the state's 10 highest low SES student percentage school districts; and 4 of the state's 10 highest EL percentage school districts.

4. The 17 consistently higher science-achievement school districts (2013–2017 mean scale scores of 203 or higher) had student populations of 10,000 to 65,000. Of these, 12 had low SES student percentages greater than 50 percent.

5. Of the 26 very small school districts (fewer than 8,000 students), 16 had low SES student percentages of 60 to 99 percent. Of these, eleven had the lowest mean scale scores (less than 200) and pass rates (17 to 40 percent) of all 67 Florida school districts. The four consistently low outlier school districts for all three student groups had student populations of fewer than 6,000 students and 60 to 99 percent low SES students.

Discussion of Findings

This discussion is organized by the topics presented in the conceptual framework. The findings are discussed as they relate to the literature reviewed for each topic. These topics are science education, science assessments, science standards, science education challenges, and science courses.

Science Education

The overall findings of this study show that, regardless of the type of science course school districts offered, less than half of all Florida 8th grade students achieved the minimum level of science proficiency, as measured by required Florida assessments. This is consistent with the 2013 to 2017 results of similar national and international tests (Desilver, 2017; FLDOE, 2017k; NCES, 2016; 2017f; Organization for Economic Cooperation and Development [OECD], 2016a). From 2013 to 2017, there was no statistically significant increase in school district mean scale scores or pass rates for either school district group. The stagnation of student achievement and the wide dispersion of school district mean scale scores and pass rates is consistent with the national and international disagreements over the definition of the term scientific literacy (Holbrook & Rannikmae, 2009), and over the need for improvement in student science achievement (Langdon et al., 2011; National Academy of Sciences [NAS], 2007; National Science Board, 2007; U.S. Department of Education, 2016b, 2016a). The findings show that, for middle grades science education, Florida school districts are not achieving the Florida Department of Education's top goal of "highest student achievement, as indicated by evidence of student learning gains at all levels" (§1008.31 Fla. Stat., 2016; FLDOE, 2014f, para. 2).

The findings of numerically increasing and overall statistically significantly higher mean pass rates for the subject specific group may be important from a school and school district leadership perspective. The 8th grade FCAT 2.0 Science/SSA overall pass rate, not the mean scale score, is a component of Florida's school and school district grading model (FLDOE, 2017b). From 2013 to 2017, the mean pass rate for the subject-specific group was numerically higher than that of the comprehensive group. During these five years, the subject-specific group mean pass rate increased, while that of the comprehensive group decreased. In 2015 and 2017, the difference reached statistical significance, with a medium effect size. By 2017, the subject-specific group mean pass rate was more than six percentage points higher than that of the comprehensive group. For the first time for either school district group, the 2017 subject-specific group mean pass rate exceeded 50 percent. School district leaders should consider this prior to making any decision to change middle school science course offerings.

Science Assessments

While there were no significant differences in mean scale scores between the two school district groups, statistically significant differences were found in the mean pass rates and mean raw scores between the two school district groups in independent samples tests. In the subject-specific group, a higher percentage of students achieved the minimum passing score or higher than in the comprehensive group. In the subject-specific group, students answered correctly more items in three subject areas than in the comprehensive group school districts. The dispersions of school district mean scale scores for all three student groups were greater for the comprehensive group. These findings suggest that the 8th grade FCAT 2.0 Science/SSA was focused on subject-specific items, and are consistent with the literature showing that standardized assessments focus on easy-to-measure, low- and moderate-complexity items rather than on items that require students to demonstrate knowledge of the integration between scientific practices and conceptual understanding (Britton & Schneider, 2010; Pellegrino et al., 2014). These findings are also consistent with the stated structure of the 2013–2017 8th grade FCAT 2.0 Science/SSA, which derived up to 90 percent of scale score points from low- to moderate-complexity items (FLDOE, 2013b, 2014d, 2015d, 2016d, 2017l).

Science Standards

The finding of wider dispersions of mean scale scores for the comprehensive group than the subject-specific group, even when controlled for demographic factors, is consistent with the research showing that comprehensive science courses vary widely in focus, sequencing, and instructional methods (Herr, 2007; International Bureau of Education [IBE]-UNESCO 2017b; Kesidou & Roseman, 2002; NRC, 2012; Ragel, 2015; Sherriff, 2014, 2015; Stengel, 1997; Tamassia & Frans, 2014). Unlike the national Next Generation Science Standards (NGSS), which specify cross-cutting concepts and connections among the science subject areas (Hoeg & Bencze, 2017), Florida's Next Generation Sunshine State Standards (NGSSS) for science provide little guidance as to the higher-level subject-area connections students are expected to achieve (Florida State University [FSU], 2017), leaving the structure and sequence of standards for comprehensive science courses open to interpretation. The wide dispersion of mean scale scores among the comprehensive group suggests there were wide variations in the interpretation and sequencing of the NGSSS for science, and in the implementation of comprehensive science courses. Florida's NGSSS for science are classified by subject area (FSU, 2017). The assessments based on these standards are subject-specific and prioritize measurable and reproducible performance for students to demonstrate progress. Eighty percent of Florida's 109 middle grades science benchmarks are Level 1 (low-complexity) and Level 2 (moderate-complexity) benchmarks (FSU, 2017). Eighty to ninety percent of the FCAT 2.0 Science/SSA items are low to moderate complexity. Labaree (2013) noted that PISA measures what is relevant but not taught; and state assessments measure what is taught but not relevant. Similarly, school districts offering comprehensive science courses may be focused on that which is relevant (higher-level thinking and connections among the science subject areas), but not assessed on the state science assessment. This increases the importance of professional development and content knowledge for science teachers and curriculum specialists (Bybee, 1997; Carr & Harris, 2001; DeBoer, 2000; Hattie, 2009; NAS, 2000; Stoica, 2015; Tyler, 1950; Zhbanova et al., 2010).

The narrower mean scale score dispersions for the subject-specific group suggest that the courses and instruction in these school districts may have been more standards-focused than in the comprehensive group. This is consistent with research showing that narrowing of the curriculum is often a consequence of standards-based education (Adler et al., 2006; Au, 2007; David, 2011; Hargrove et al., 2000; Vogler & Virtue, 2007).

Science Education Challenges

The findings of this study support the research showing that large school districts face unique challenges (Amah et al., 2013; Borman & Kimball, 2005; Boser, 2013; Bouck, 2004; Davis, 2003; Driscoll et al., 2003; Hannaway & Kimball, 1998; Johnson et al., 2004; Koran, 2016). Student populations among Florida's 67 public school districts were highly positively skewed. Only seven school districts had student populations of over 100,000 students. The median student population was just under 13,000 students. Yet, regardless of school district or student group, the low science-achievement school districts had greater than average mean student populations and included 4 of the state's 10 largest school districts (including the two largest). A statistically significant, medium, negative correlation was found between school district student population and school district mean scale scores for low SES students for both school district groups. However, large school district size has been shown to have a negative impact on overall student achievement, even after controlling for the concentration of low SES students (Amah et al., 2013; Driscoll et al., 2003; Lippman et al., 1996).

The research suggests that the negative correlation between school district size and student achievement may be due to many factors, including limited responsiveness to the needs of students and parents, less agility in adapting to changing student enrollment, less supportiveness of teachers, and less control over implementation of new curricula (Bouck, 2004; Driscoll et al., 2003; Johnson et al., 2004; Koran, 2016; McLaughlin, 2014; Rivera Maulucci, 2010; Rivera Maulucci et al., 2014). Driscoll et al (2003) showed that population and poverty density, not just school district size, are the primary drivers of larger achievement gaps in large school districts.

Low achievement extends to very small school districts as well (Driscoll et al., 2003). This is supported by the finding that school districts with the lowest mean scales scores and pass rates had populations less than 8,000 students and low SES student percentages of 60 to 99 percent. The four consistently low outlier school districts for all three student groups had fewer than 6,000 students and low SES student percentages of 60 to 99 percent.

The findings of this study support the research on the educational challenges for students in poverty. For both school district groups, the achievement gap between all students and low SES students was statistically significant, with a very large effect size, according to the criteria of Sawilowsky (2003). The percentage of low SES students is normally distributed among Florida's 67 public school districts. For all student groups in both school district groups, the 19 consistently low science-achievement school districts had higher than average percentages, or densities, of low SES students. The low science-achievement school districts included 9 of the state's 10 highest poverty (by low SES percentage) school districts.

Together, these findings support research on the persistent achievement gap for low SES students (Becker & Luthar, 2002; Hanushek, 2010; Hattie, 2009; Ladd, 2012; Miller et al., 2013; NCES, 2017a; OECD, 2016c; Wiseman, 2012). The findings are also consistent with research showing that the low SES student achievement gap is not unique to large, urban school districts , Wiseman (Bouck, 2004; Wiseman, 2012), and that poverty is the most significant factor in explaining educational achievement gaps (Lara-Cinisomo et al., 2004).

The findings of this study are also consistent with research on the educational challenges of English learners (ELs). The percentage of ELs is positively skewed among Florida's 67 public school districts. This skew helps explain why no statistically significant correlation was found between school district EL percentage and school district mean scale scores. Forty of the state's 67 school districts had EL percentages of less than five percent. Twenty-five school districts had EL percentages from 5 to 15 percent.

The two highest EL percentage school districts had 18 and 20 percent ELs. One of these was the state's largest school district with more than 300,000 students, and the state's highest percentage of low SES students (72 percent). The other was a medium-sized school district (about 60,000 students), with the state's second-highest percentage of low SES students (about 66 percent). These findings are consistent with the research showing that ELs often also are socioeconomically disadvantaged (Cosentino de Cohen et al., 2005).

For both school district groups, the achievement gap between all students and ELs was statistically significant, with a huge effect size, according to the criteria of Sawilowsky (2003).

The consistently low science-achievement school districts included 4 of the state's 10 highest EL percentage school districts. These findings suggest that science teachers may not have received the special professional development needed to help ELs learn scientific concepts and processes while their language skills are developing (Amaral et al., 2002; Fathman et al., 1992; Lara-Alecio et al., 2012; Santau et al., 2010).

The differences in results between independent samples tests and paired samples tests support the importance of using demographically similar school districts when making comparisons of student achievement (Becker & Luthar, 2002; Hanushek, 2010; Hattie, 2009; Ladd, 2012; Miller et al., 2013; Wiseman, 2012). Overall, three of the five independent samples *t*-tests of school district mean scale scores approached statistical significance, with the subject-specific group being higher. Paired samples *t*-tests and non-parametric tests did not approach significance. Independent samples *t*-tests of school district mean pass rates showed statistically significantly higher mean pass rates for the subject-specific group in two years, but paired samples *t*-tests found no statistically significant differences.

Science Courses

Multiple levels of quantitative analyses found no statistically significant differences in the 2013–2017 8th grade FCAT 2.0 Science/SSA school district mean scale scores for the three groups of students between the two school district groups. No statistically significant differences, increases, or decreases were found in school district mean scale scores or pass rates from 2013 to 2017. At face value, these results support the findings of the few similar studies that compared student science achievement of students in comprehensive and subject-specific science courses (Alwardt, 2011; Åström, 2007, 2008; Åström & Karlsson, 2012; Clifford, 2016;

Faulkner, 2012; Hattie, 2009, 2017; Tamassia & Frans, 2014). However, the statistically significant differences in mean pass rates and raw scores between the independent samples of the two school district groups show that subject-specific courses may help a higher percentage of students reach the minimum level of achievement.

The dispersions of the comprehensive group school district mean scale scores and mean pass rates were greater each year than that of the subject-specific group. Many of the differences were found to be statistically significant differences in variance. School districts in the comprehensive group consistently had both higher and lower mean scale scores than those in the subject-specific group. This was evident for all three groups of students. These differences in dispersion were present in both independent school district samples with different sample sizes, and paired school district samples with equal numbers of school districts from either group, controlled for student population, percentage of low SES students, and percentage of ELs.

The consistently greater dispersion of mean scale scores for the comprehensive group has two implications. One implication may be that course structure and implementation was less consistent among school districts in the comprehensive group than among those in the subject-specific group. This would be consistent with research that found widely varying terms, definitions, course and curriculum structures, and methods of implementation for comprehensive science curricula (Herr, 2007; IBE-UNESCO, 2017b; Kesidou & Roseman, 2002; NRC, 2012; Ragel, 2015; Sherriff, 2014, 2015; Stengel, 1997; Tamassia & Frans, 2014). Some school districts may have used comprehensive science curricular materials from national publishers that attempt to accommodate the varying standards of as many states as possible (Kesidou & Roseman, 2002). These curricula were found to be largely comprised of previously-written, subject-specific curricula, re-arranged into comprehensive curricula with no true integration among the subject-specific concepts (Kesidou & Roseman, 2002). Another possible implication is that there may have been consist course structure and implementation, but its effects on student science achievement varied among groups of students in Florida's economically and culturally diverse student population. This would be consistent with research showing the differing effectiveness of science instructional methods among students with various cultural and economic backgrounds (R. D. Anderson & Helms, 2001; Gao, 2014; Gao & Wang, 2016; Tobin, 1986).

Of the 19 consistently low-science achievement school districts, 95 percent were in the comprehensive group. Of all 67 Florida school districts, 75 percent were in the comprehensive group. While other factors may have contributed to the low science achievement in these school districts, the challenges of inconsistent definition, structure, and implementation of the comprehensive courses may have been involved as well (Herr, 2007; IBE-UNESCO, 2017b; Kesidou & Roseman, 2002; NRC, 2012; Ragel, 2015; Stengel, 1997; Tamassia & Frans, 2014). Florida's science standards are subject-area focused, and provide little guidance for the implementation of comprehensive science courses (FLDOE, 2017l, 2017n, 2017o).

From 2013 to 2017, the subject-specific group school district mean pass rate for all students increased, while that of the comprehensive group decreased. In 2015 and 2017, independent samples tests showed the school district mean pass rates for all students were statistically significantly higher in the subject-specific group than in the comprehensive group. In 2013, the school district mean pass rates for all students in each group were almost identical. By 2017, the subject-specific group school district mean pass rate for all students increased by almost four percentage points and was more than six percentage points higher than that of the comprehensive group. Additionally, during this time, the dispersion of the school district mean pass rate for all students decreased in the subject-specific group, while it increased in the comprehensive group.

These differences in pass rates may seem to contradict the findings of no statistically significant in mean scale scores. However, this dichotomy is explained by the differences in the dispersions of both measures of student achievement. From 2013 to 2017, the dispersions of both school district mean scale scores and pass rates increased for the comprehensive group and decreased for the subject-specific group. These changes in dispersion were bilateral and offsetting for both groups, leaving the overall mean scale scores unaffected. However, there was just enough change around the minimum passing score to affect the mean pass rates. This further supports the research showing that science standards, and the assessments based on them, are subject-area focused. While the standards provide little guidance for the implementation of comprehensive science courses, they provide clearer guidance for the consistent implementation of subject-specific courses (FLDOE, 2017l, 2017n, 2017o; Hoeg & Bencze, 2017; Marx & Harris, 2006; Vogler & Virtue, 2007).

This study found that the school district mean raw scores for all students were statistically significantly higher for the subject-specific group than for the comprehensive group in the subject areas of nature of science, life science, and physical science. Life science and physical science are typically offered in the 7th and 8th grades in subject-specific school districts (FLDOE, 2015a; FSU, 2017). Nature of science standards are included in 6th, 7th, and 8th grades in both comprehensive and subject-specific courses (FLDOE, 2015a; FSU, 2017). This finding suggests that the learning science principle of recency may have helped students in the subject-specific group answer correctly more items requiring knowledge learned most recently (G. H. Bower & Hilgard, 1981; Kahana et al., 2008; Roediger, 2006). The learning science principles of exercise and multiple examples may have helped students in this group answer correctly more items requiring knowledge of the nature of science (G. H. Bower & Hilgard, 1981; Kahana et al., 2008). However, the raw score differences may have been caused by differences in the various IRT-based assessment forms (de Ayala, 2008; Kim, 2007; Orr, 2008; Rich, 2017; Visone, 2009, 2010).

Implications for Practice

1. The overall pass rate for the subject-specific group grew to more than six percentage points higher than that of the comprehensive group from 2013 to 2017. The 8th grade FCAT 2.0 Science/SSA school and school district mean pass rates, not mean scale scores, are a component of Florida's school grade model (FLDOE, 2017b). The mean pass rate for the subject-specific group increased from 2013 to 2017 and in independent samples tests, was statistically significantly higher overall than that of the comprehensive group in 2015 and 2017. For the first time since before 2013 for either group of school districts, the subject-specific group mean pass rate exceeded 50 percent in 2017. While these differences in pass rates were not statistically significant when demographically similar school districts were compared, leaders of lower science-achievement school districts in both groups would be wise to study more closely demographically-similar, higher-achievement school districts in the subject-specific group prior to making any changes to science course offerings.

2. Demographics are important in making comparisons of science achievement among Florida's school districts. Different groups of students may respond differently to different types of science courses. While there were no statistically significant differences, overall the subject-specific group had consistently higher mean scale scores for all students and for ELs, and the comprehensive group had consistently higher mean scale scores for low SES students. When demographically similar school districts were compared, the comprehensive group had numerically higher mean scale scores for all three student groups.

3. Very large and very small school districts are at a disadvantage, regardless of the type of science course offering. Both the largest and smallest school districts were consistently among the lowest in science achievement. While these school districts generally had high percentages of low SES students, poverty alone did not explain the low achievement. Thirteen of the state's 17 consistently higher science-achievement school districts had similar low SES student percentages (greater than 50 percent). However, these school districts had medium-sized populations of 8,000 to 62,000 students.

4. Some school districts in the comprehensive group achieved higher mean scale scores despite high concentrations of poverty. While high concentrations of poverty were found to be correlated with lower student achievement for low SES students, seven of the state's 17 consistently high-science-achievement school districts offered comprehensive science courses and had low SES student percentages greater than fifty percent. Leaders of lower science-achievement school districts in both groups would be wise to study these seven school districts to determine how this higher level of student success was achieved.

5. Comprehensive science courses may afford school district leaders more flexibility in teacher assignments. Low SES students tended to have lower achievement in larger school districts, and most of these larger school districts offered comprehensive courses. This may seem to be an implication against comprehensive science courses. However, these school districts may have offered comprehensive science courses due to the flexibility it affords with regard to teacher assignments. Changing teacher assignments is more often necessary in larger than in smaller school districts (Driscoll et al., 2003; Koran, 2016). Teachers with middle grades general science licensure, or subject-specific licensure in all three subject areas (physical, life, and Earth/space science), may be reassigned among the three middle grades more easily than those with only a single subject-specific licensure. In this respect, comprehensive science may offer educational leaders of larger school districts more flexibility in responding to changing needs than subject-specific science. However, this same flexibility may be achieved in subject-specific school districts by requiring teachers to have licensure in middle grades general science or in all three subject areas.

6. Comprehensive science courses are not well-defined in the Florida NGSSS science standards. This is evident in the wide dispersion of mean scale scores and pass rates for school districts in the comprehensive group. Science teacher preparation and continuing professional development is necessary for both traditional pre-service teachers and for the growing force of alternatively certified science teachers entering the teaching profession with varying levels of experience and subject matter education. This need extends to science curriculum specialists who develop each school district's comprehensive science courses. School district leaders considering a change to comprehensive science course offerings would be wise to consult with

leaders of higher achievement school districts with similar demographics prior to making any change.

7. School district leaders should not expect to effect improvement in 8th grade Statewide Science Assessment mean scale scores or pass rates simply by changing the type of middle grades science course offering. School district leaders in either group of school districts may better impact student science achievement not by changing the science curriculum, but by focusing on building positive relationships with students, quality of instruction, and teacher preparation and professional development in the instructional methods that lead to higher science achievement of a diverse student population (Bybee, 1995; Drits-Esser & Stark, 2015; Gao, 2014; Gao & Wang, 2016; Hattie, 2009, 2017; Pringle et al., 2017; Sturges, 1976; Waters et al., 2003).

8. Florida's NGSSS for science and the FCAT 2.0 Science/SSA are subject-specific. If scientific literacy, higher-level learning, and greater student achievement are indeed among the state's top educational goals, science standards and benchmarks should be revised to include standards and benchmarks that cross subject-area boundaries. The results of standardized science assessments are used, in part, to evaluate educational leaders, science curriculum specialists, science teachers, as well as science students. The standards and benchmarks on which the assessments are based should clarify the higher-level connections students are expected to master.

9. Science educators should be able to use the Statewide Science Assessment results to inform improvements in instruction and curriculum. The 8th grade FCAT 2.0 Science/SSA is standards-based, expertly-written, well-validated, well-maintained, and rich with useful but

unobtainable information (FLDOE, 2013b, 2014c, 2014d, 2015d, 2016d). The wide dispersion of mean scale scores and pass rates among school districts in the comprehensive group, and the overall stagnation in mean scale scores and pass rates among both groups of school districts, implies that some school districts are struggling to find a science curriculum that supports student learning. To increase student science achievement, science educators in both groups need detailed assessment data that clearly shows the standards and benchmarks in which their students are lacking. State-level educational leaders should consider including a teacher questionnaire with every administration of the SSA, similar to those used in PISA, TIMSS, and NAEP science assessments (National Center for Education Statistics, 2015b, 2015d, 2015e). A questionnaire such as this would facilitate analysis of data concerning instructional methods and their effects on science achievement among Florida's economically and culturally diverse student population.

Recommendations for Further Research

This study is limited and delimited in several respects, presenting opportunities for further study. This study was delimited to the use of only archival data, without observation or other data collection regarding individual classroom instructional approaches, science teacher experience, teacher turnover, or student reading or math achievement levels. Seven of the state's seventeen school districts with mean scale scores above the minimum passing score offered comprehensive science courses. These school districts exceeded the maximum score of all subject-specific school districts yet had greater than 50 percent low SES students. These school districts should be studied, both qualitatively and quantitatively, to determine how this higher student success was achieved despite the demographic factors that, statistically, should have led to low science achievement. Published studies that analyzed different science instructional methods and their effects on student achievement on international tests (Gao, 2014; Gao & Wang, 2014, 2016) may be used as models. Observations of the instructional methods, quality and implementation of instruction, professional development, student cultural backgrounds, and teacher-student relationships, in school districts offering either type of science course, may help lower science-achievement school districts in both groups. Interviews with school district leaders and science curriculum specialists about the reasons for their decisions to change science curricula, and their experiences in doing so, may be helpful to other school district leaders considering a change in science course offerings.

This study was delimited to analysis of only Florida 8th grade FCAT 2.0 Science/SSA scores. Further research of similar data for other state, national, or international science assessments would add to the scant body of literature regarding comprehensive vs. subject-specific science education. Prior to this study, only seven related studies were found. Of these, only two were published and peer-reviewed. Both were conducted Europe. Four, like this study, were doctoral dissertations, and one, a master's thesis (Alwardt, 2011; Åström, 2007, 2008; Åström & Karlsson, 2012; Clifford, 2016; Faulkner, 2012; Hattie, 2009, 2017; Tamassia & Frans, 2014).

This study focused on student science achievement only. The perspectives of school and school and school district leaders was not considered. Interviews with school and school district leaders that have made, or are considering making a change, in the type of science course offering, may provide valuable insight for other school and school district leaders. These interviews might include questions about how the need for flexibility in teacher assignments and the middle school acceleration component of Florida's school grading system impacted their decisions (FLDOE, 2017a, 2017m).

This study was delimited to consideration of only comprehensive and subject-specific middle school science courses that address Florida's NGSSS for science for grades 6 through 8. This study did not consider the student achievement impacts of middle grades science research elective courses offered in addition to the normal science courses, nor the high school biology courses offered in lieu of the normal middle grades science courses in Florida's middle school acceleration program (FLDOE, 2017a, 2017m). A comparison of student science achievement in school districts that offered these courses to those that did not may provide information useful to school district leaders in the search for options to improve student science achievement and bolster school and school district grades.

Conclusions

As with the 8th grade student achievement on similar national and international science assessments, Florida's 8th grade student achievement on the 2013–2017 8th grade Florida Comprehensive Assessment Test for Science (FCAT 2.0 Science)/Statewide Science Assessment (SSA) was stagnant. For those five years, half or more students did not achieve the minimum passing score. Up to two thirds of students in many school districts with large student populations, high percentages of low socioeconomic status students, and high percentages of English learners did not achieve the minimum passing score. There is disagreement as to whether this constitutes a national emergency or a simple lack of scientific literacy among students. However, this study shows that the number one goal of the FLDOE of "highest student achievement" (FLDOE, 2014f, para. 2) is not being met in the area of middle grades science.

To improve evaluation outcomes, many Florida school district leaders have followed a national trend in changing middle grades science curriculum from subject-specific, discipline-based, layered, field-specific, or traditional science courses to comprehensive, integrated, spiral, interdisciplinary, multidisciplinary, thematic, or general science courses. However, there was a lack of research showing if either type of course improved student achievement on standardized science assessments. Using multiple levels of quantitative tests, this study compared the 2013–2017 8th grade FCAT 2.0 Science/SSA school district mean scale scores for three groups of students in two groups of school districts: those that offered comprehensive science courses, and those that offered subject-specific science courses. The three groups of students were all 8th grade students assessed, the low socio-economic status (SES) student subgroup, and the English learners (ELs) subgroup.

This study is significant in that it helps fill a void in science education research regarding differences in student science achievement between school districts offering comprehensive and subject-specific science courses and provides evidence for Florida school district leaders to use in making informed decisions regarding changes in middle grades science course offerings and curricula. Controlling for school district demographic factors of student population, percentage of low SES students, and percentage of ELs, no statistically significant differences were found in school district and school-level mean scale scores or pass rates between the two groups of school districts. These results were consistent with the few similar, prior studies of differences in student achievement between groups of students in comprehensive and subject-specific science courses.

The results of this study were consistent with research outside the area of science education. Tests and analyses showed a significant, negative correlation between the school district mean scale score for low SES students and school district student population, regardless of the type of science course offered. In both school district groups, those with the lowest student achievement generally had greater overall student populations, percentages of low SES students, and percentages of ELs than higher-achievement school districts. However, very small school districts with high percentages of low SES students were also consistently among the lowest science-achievement school districts.

The findings of this study show that, if properly implemented, comprehensive science courses may lead to higher student achievement than subject-specific courses. The maximum school district mean scale scores for all student groups were consistently achieved in school districts that offered comprehensive science courses. However, school districts that offered comprehensive science courses also had the lowest mean scale scores for all student groups, and consistently wider dispersions of mean scales scores and pass rates. This suggests that either there was less consistency in course structure and implementation among these school districts, or that there was consistency, but its effects varied widely among the diverse student populations.

The primary implication of this study is that school district educational leaders should not expect to effect significant improvement in student achievement on the 8^{th} grade FCAT 2.0 Science/SSA simply by changing the type of science course offering. High science-achievement school districts in both groups should be studied more closely to determine how they are achieving this success, particularly those few that have achieved higher success despite high concentrations of students in poverty. A limitation of this study is that it did not analyze the differences in instructional methods employed by science teachers of either comprehensive or subject-specific science teachers. The wide dispersion of mean scale scores and pass rates among school district in the comprehensive group suggests that further study is needed to determine which science courses and instructional methods work best for each group of students among Florida's diverse student population. Change is needed to break the stagnation in student science achievement, improve scientific literacy for all students, and prepare Florida's students for future STEM careers. Any changes to middle grades science courses should be done with consideration to the needs of students from diverse cultural backgrounds, teacher professional development and support, course structure and alignment to standards, school district demographics, and the need for flexibility in teacher assignments.

APPENDIX A: INSTITUTIONAL REVIEW BOARD EXEMPTION LETTER

University of Central Florida Institutional Review Board
Office of Research & Commercialization
12201 Research Parkway, Suite 501
Orlando, Florida 32826-3246
Telephone: 407-823-2901, 407-882-2012 or 407-882-2276
www.research.ucf.edu/compliance/irb.html

NOT HUMAN RESEARCH DETERMINATION

From : **UCF Institutional Review Board #1**
FWA00000351, IRB00001138

To : **Kenneth R. Moore**

Date : **May 03, 2017**

Dear Researcher:

On 05/03/2017 the IRB determined that the following proposed activity is not human research as defined by DHHS regulations at 45 CFR 46 or FDA regulations at 21 CFR 50/56:

Type of Review:	Not Human Research Determination
Project Title:	A COMPARISON OF FLORIDA 8TH GRADE STATEWIDE SCIENCE ASSESSMENT MEAN SCALE SCORES 2013 THROUGH 2016 FOR SCHOOL DISTRICTS USING SUBJECT-FOCUSED VS. COMPREHENSIVE MIDDLE SCHOOL SCIENCE CURRICULA
Investigator:	Kenneth R. Moore
IRB ID:	SBE-17-13142
Funding Agency:	
Grant Title:	
Research ID:	NA

University of Central Florida IRB review and approval is not required. This determination applies only to the activities described in the IRB submission and does not apply should any changes be made. If changes are to be made and there are questions about whether these activities are research involving human subjects, please contact the IRB office to discuss the proposed changes.

On behalf of Sophia Dziegielewski, Ph.D., L.C.S.W., UCF IRB Chair, this letter is signed by:

Signature applied by Gillian Amy Mary Morien on 05/03/2017 09:09:18 AM EDT

IRB Coordinator

APPENDIX B: PERMISSIONS FROM COPYRIGHT HOLDERS

Kenneth R. Moore
5776 Marble Ct.
Winter Park FL 32792

15 June 2017

Eric R. Banilower
Vice President, Horizon Research, Inc.
326 Cloister Court
Chapel Hill, NC 27514

Dear Mr. Banilower,

This letter is to confirm the recent email from Mr. Whiffen. I am completing a doctoral dissertation at the University of Central Florida entitled " A Comparison of Florida 8th Grade Statewide Science Assessment Mean Scale Scores 2013 Through 2017 for School Districts Using Subject-Focused vs. Comprehensive Middle School Science Curricula". I would like your permission to reprint in my dissertation a table from Horizon Research, Inc. 2012 National Survey of Science and Mathematics Education. Specifically, Table 4.3 on page 54, "Type of Middle School Science Courses Offered, by Grade".

Type of Middle School Science Courses Offered, by Grade

	Percent of Schools		
	Grade 6	Grade 7	Grade 8
Single-Discipline Science Courses Only	36 (3.6)	46 (3.8)	47 (3.8)
Coordinated or Integrated Science Courses Only	45 (4.1)	38 (3.7)	36 (3.7)
Both	19 (3.5)	15 (3.6)	18 (3.5)

Banilower, E. R., Smith, P. S., Weiss, I. R., Malzahn, K. A., Campbell, K. M., & Weis, A. M. (2013). *Report of the 2012 national survey of science and mathematics education.* Horizon Research, Inc.

The requested permission extends to any future revisions and editions of my thesis/dissertation, including non-exclusive world rights in all languages. These rights will in no way restrict republication of the material in any other form by you or by others authorized by you. Your signing of this letter will also confirm that you own or your company owns the copyright to the above-described material.

If these arrangements meet with your approval, please sign this letter below and return it to me in the enclosed return envelope. Thank you for your attention in this matter.

Sincerely,

Kenneth R. Moore

PERMISSION GRANTED FOR THE USE REQUESTED ABOVE.

By: ______________________ Date: 6/22/17

Eric R. Banilower
Vice President, Horizon Research, Inc.

OECD General Permission for Use and Adaptation of PISA Tables and Figures.

APPENDIX C: SCIENCE COURSE OFFERINGS BY SCHOOL DISTRICT

Table 88

Science Course Offerings by School District, 2012-13 through 2015-16 (*N* = 67 School Districts)

School district	Year				
	2013[a]	2014[b]	2015[c]	2016[d]	2017[e]
Abel	Comprehensive	Comprehensive	Comprehensive	Comprehensive	Comprehensive
Abioye	Comprehensive	Comprehensive	Comprehensive	Comprehensive	Comprehensive
Adil	Subject-specific	Subject-specific	Subject-specific	Subject-specific	Comprehensive
Aishah	Comprehensive	Comprehensive	Comprehensive	Comprehensive	Comprehensive
Amias	Comprehensive	Comprehensive	Comprehensive	Comprehensive	Comprehensive
Andreas	Comprehensive	Comprehensive	Comprehensive	Comprehensive	Comprehensive
Antigone	Comprehensive	Comprehensive	Comprehensive	Comprehensive	Comprehensive
Barnaby	Subject-specific	Subject-specific	Subject-specific	Subject-specific	Subject-specific
Benson	Comprehensive	Comprehensive	Comprehensive	Comprehensive	Comprehensive
Blythe	Comprehensive	Comprehensive	Comprehensive	Comprehensive	Comprehensive
Branch	Subject-specific	Subject-specific	Comprehensive	Comprehensive	Comprehensive
Brande	Comprehensive	Comprehensive	Comprehensive	Comprehensive	Comprehensive
Callahan	Comprehensive	Comprehensive	Comprehensive	Comprehensive	Comprehensive
Ciara	Comprehensive	Comprehensive	Comprehensive	Comprehensive	Comprehensive
Clifford	Comprehensive	Comprehensive	Comprehensive	Comprehensive	Comprehensive
Cornell	Subject-specific	Subject-specific	Subject-specific	Subject-specific	Subject-specific
Desta	Comprehensive	Comprehensive	Comprehensive	Comprehensive	Comprehensive
Earline	Subject-specific	Comprehensive	Comprehensive	Comprehensive	Comprehensive
Emmett	Comprehensive	Comprehensive	Comprehensive	Subject-specific	Subject-specific
Evander	Comprehensive	Comprehensive	Comprehensive	Comprehensive	Comprehensive
Everett	Comprehensive	Comprehensive	Comprehensive	Comprehensive	Subject-specific
Farida	Comprehensive	Comprehensive	Comprehensive	Comprehensive	Comprehensive
Firdaus	Comprehensive	Comprehensive	Comprehensive	Comprehensive	Comprehensive

School district	Year				
	2013[a]	2014[b]	2015[c]	2016[d]	2017[e]
Gereon	Comprehensive	Comprehensive	Comprehensive	Comprehensive	Comprehensive
Gilad	Comprehensive	Comprehensive	Comprehensive	Comprehensive	Comprehensive
Giselle	Subject-specific	Subject-specific	Subject-specific	Subject-specific	Subject-specific
Gottfried	Subject-specific	Comprehensive	Comprehensive	Comprehensive	Comprehensive
Husniya	Comprehensive	Comprehensive	Comprehensive	Comprehensive	Comprehensive
Judith	Comprehensive	Comprehensive	Comprehensive	Comprehensive	Comprehensive
Junayd	Comprehensive	Comprehensive	Comprehensive	Comprehensive	Comprehensive
Katharine	Comprehensive	Comprehensive	Comprehensive	Comprehensive	Subject-specific
Katlyn	Subject-specific	Subject-specific	Subject-specific	Subject-specific	Subject-specific
Kimberly	Comprehensive	Comprehensive	Comprehensive	Comprehensive	Comprehensive
Lavender	Comprehensive	Comprehensive	Comprehensive	Comprehensive	Comprehensive
Lawson	Subject-specific	Subject-specific	Subject-specific	Subject-specific	Subject-specific
Linwood	Subject-specific	Subject-specific	Subject-specific	Subject-specific	Subject-specific
Lucas	Comprehensive	Comprehensive	Comprehensive	Comprehensive	Subject-specific
Macy	Comprehensive	Comprehensive	Comprehensive	Comprehensive	Comprehensive
Madeline	Comprehensive	Comprehensive	Comprehensive	Comprehensive	Comprehensive
Manfred	Comprehensive	Comprehensive	Comprehensive	Comprehensive	Comprehensive
Mirela	Comprehensive	Comprehensive	Comprehensive	Subject-specific	Subject-specific
Odell	Comprehensive	Comprehensive	Comprehensive	Comprehensive	Comprehensive
Oneida	Comprehensive	Comprehensive	Comprehensive	Comprehensive	Comprehensive
Phokas	Comprehensive	Comprehensive	Comprehensive	Comprehensive	Comprehensive
Placido	Subject-specific	Subject-specific	Subject-specific	Subject-specific	Comprehensive
Renato	Subject-specific	Subject-specific	Subject-specific	Subject-specific	Subject-specific
Reuben	Comprehensive	Comprehensive	Comprehensive	Comprehensive	Comprehensive
Rilla	Comprehensive	Comprehensive	Comprehensive	Comprehensive	Comprehensive
Roel	Comprehensive	Comprehensive	Comprehensive	Comprehensive	Comprehensive
Ross	Comprehensive	Comprehensive	Comprehensive	Comprehensive	Comprehensive

School district	Year 2013[a]	2014[b]	2015[c]	2016[d]	2017[e]
Rukiye	Subject-specific	Subject-specific	Subject-specific	Subject-specific	Subject-specific
Sabela	Comprehensive	Comprehensive	Comprehensive	Comprehensive	Comprehensive
Samson	Subject-specific	Subject-specific	Subject-specific	Subject-specific	Subject-specific
Shanna	Comprehensive	Comprehensive	Comprehensive	Comprehensive	Subject-specific
Suhaila	Comprehensive	Comprehensive	Comprehensive	Comprehensive	Comprehensive
Sulayman	Subject-specific	Subject-specific	Subject-specific	Subject-specific	Comprehensive
Sunita	Comprehensive	Comprehensive	Comprehensive	Comprehensive	Comprehensive
Tarek	Subject-specific	Comprehensive	Comprehensive	Comprehensive	Subject-specific
Thorsten	Comprehensive	Comprehensive	Comprehensive	Comprehensive	Comprehensive
Tirtzah	Comprehensive	Comprehensive	Comprehensive	Comprehensive	Comprehensive
Viktor	Subject-specific	Subject-specific	Subject-specific	Subject-specific	Subject-specific
Walker	Subject-specific	Subject-specific	Subject-specific	Subject-specific	Comprehensive
Westley	Comprehensive	Comprehensive	Comprehensive	Comprehensive	Comprehensive
Xaver	Comprehensive	Comprehensive	Comprehensive	Comprehensive	Comprehensive
Yishma	Comprehensive	Comprehensive	Comprehensive	Comprehensive	Comprehensive
Zaina	Comprehensive	Comprehensive	Comprehensive	Comprehensive	Comprehensive
Zakiah	Comprehensive	Comprehensive	Comprehensive	Comprehensive	Comprehensive
Total Comprehensive	49	52	53	51	50
Total Subject-specific	18	15	14	16	17
Totals	67	67	67	67	67

Notes: Fictitious school district names used. Data compiled from FLDOE Pk-12 Public School Data Publications and Reports. [a]Course Enrollment Survey 3, 2012-13 (FLDOE, 2013a). [b]Course Enrollment Survey 3, 2013-14 (FLDOE, 2014b). [c]Course Enrollment Survey 3, 2014-15 (FLDOE, 2015b). [d]Course Enrollment Survey 3, 2015-16 (FLDOE, 2016b). [e] Course Enrollment Survey 3, 2015-16 (FLDOE, 2017f)

APPENDIX D: SCHOOL DISTRICT PAIRED SAMPLES

Table 89

Research Questions 2 and 3 School District Paired Sample, 2013 (*N* = 18 school district pairs)

Pair number	School district	Course type	Student population[a]	Low SES (%)[b]	EL (%)[c]
1	Zaina	Comprehensive	2,647	79.7	0.3
1	Tarek	Subject-specific	1,040	70.0	1.4
2	Emmett	Comprehensive	1,193	60.4	5.9
2	Katlyn	Subject-specific	5,995	67.7	4.3
3	Lucas	Comprehensive	4,982	56.8	0.0
3	Earline	Subject-specific	1,930	56.5	0.0
4	Clifford	Comprehensive	6,920	63.9	0.8
4	Cornell	Subject-specific	9,797	65.1	0.9
5	Roel	Comprehensive	7,811	52.5	3.2
5	Lawson	Subject-specific	7,990	59.7	3.6
6	Phokas	Comprehensive	11,076	45.2	0.9
6	Viktor	Subject-specific	8,358	44.1	7.6
7	Lavender	Comprehensive	26,634	57.2	1.8
7	Gottfried	Subject-specific	12,920	55.9	2.2
8	Rilla	Comprehensive	15,307	63.1	0.9
8	Samson	Subject-specific	16,355	62.5	1.5
9	Abel	Comprehensive	18,011	56.6	6.2
9	Adil	Subject-specific	41,096	52.1	5.9
10	Suhaila	Comprehensive	33,432	44.5	1.8
10	Linwood	Subject-specific	27,826	49.0	1.8
11	Kimberly	Comprehensive	35,244	36.0	1.6
11	Renato	Subject-specific	29,786	39.7	2.4
12	Farida	Comprehensive	40,670	61.1	1.2
12	Walker	Subject-specific	41,990	67.1	4.8
13	Ciara	Comprehensive	67,153	55.1	4.1
13	Placido	Subject-specific	41,495	57.3	4.1
14	Everett	Comprehensive	46,165	55.2	9.6
14	Barnaby	Subject-specific	43,789	61.2	14.2
15	Ross	Comprehensive	71,228	45.4	3.2
15	Sulayman	Subject-specific	64,463	44.8	3.6
16	Abioye	Comprehensive	85,765	65.3	7.7
16	Branch	Subject-specific	96,937	66.6	10.6
17	Sunita	Comprehensive	179,514	54.7	11.3
17	Giselle	Subject-specific	260,226	56.9	9.9
18	Husniya	Comprehensive	200,466	57.5	12.3
18	Rukiye	Subject-specific	183,066	62.1	13.5

Notes: Fictitious school district names are used. [a]Course Enrollment by School, Survey 3, 2012-13 (FLDOE, 2013a). [b]Student Enrollment by District 2012-17 (FLDOE, 2017i). [c]Economic Status by District 2012-17 (FLDOE, 2017i). [d]ELL Students by District 2012-17 (FLDOE, 2017i).

Table 90

Research Questions 2 and 3 School District Paired Sample, 2014 (*N* = 15 school district pairs)

Pair number	School district	Course type[a]	Student population[b]	Low SES (%)[c]	EL (%)[d]
1	Desta	Comprehensive	5,508	42.8	3.9
1	Viktor	Subject-specific	8,453	45.8	7.8
2	Roel	Comprehensive	7,977	53.1	3.1
2	Lawson	Subject-specific	8,228	58.8	3.0
3	Manfred	Comprehensive	5,946	56.0	5.6
3	Katlyn	Subject-specific	5,967	51.6	4.4
4	Clifford	Comprehensive	6,802	66.3	0.9
4	Cornell	Subject-specific	10,072	62.9	0.9
5	Lavender	Comprehensive	26,850	58.0	1.8
5	Linwood	Subject-specific	28,001	52.8	1.9
6	Rilla	Comprehensive	15,008	63.3	0.9
6	Samson	Subject-specific	16,201	61.0	1.8
7	Kimberly	Comprehensive	35,487	41.1	1.6
7	Renato	Subject-specific	30,026	41.7	2.6
8	Suhaila	Comprehensive	33,514	39.9	1.9
8	Adil	Subject-specific	41,277	44.8	6.1
9	Mirela	Comprehensive	58,175	65.4	17.9
9	Barnaby	Subject-specific	44,747	61.9	14.4
10	Ciara	Comprehensive	67,779	52.7	4.4
10	Placido	Subject-specific	41,591	57.9	4.6
11	Everett	Comprehensive	46,788	58.3	9.9
11	Branch	Subject-specific	97,972	59.8	10.8
12	Junayd	Comprehensive	60,908	61.8	5.5
12	Walker	Subject-specific	42,017	54.7	5.3
13	Ross	Comprehensive	71,024	44.7	3.3
13	Sulayman	Subject-specific	64,891	45.6	3.9
14	Sunita	Comprehensive	183,452	59.7	11.2
14	Giselle	Subject-specific	262,755	59.9	10.2
15	Husniya	Comprehensive	203,988	60.0	12.2
15	Rukiye	Subject-specific	187,274	61.1	13.1

Notes: Fictitious school district names are used. [a]Course Enrollment by School, Survey 3, 2013-14 (FLDOE, 2014b). [b]Student Enrollment by District 2012-17 (FLDOE, 2017i). [c]Economic Status by District 2012-17 (FLDOE, 2017i). [d]ELL Students by District 2012-17 (FLDOE, 2017i).

Table 91

Research Questions 2 and 3 School District Paired Sample, 2015 (*N* = 15 school district pairs)

Pair number	School District	Course type[a]	Student population[b]	Low SES (%)[c]	EL (%)[d]
1	Amias	Comprehensive	2,613	49.7	2.4
1	Katlyn	Subject-specific	6,061	47.7	4.6
2	Odell	Comprehensive	1,730	47.7	9.4
2	Viktor	Subject-specific	8,457	48.6	8.5
3	Lucas	Comprehensive	4,968	57.9	0.2
3	Lawson	Subject-specific	8,294	59.9	2.8
4	Clifford	Comprehensive	6,807	67.5	1.0
4	Cornell	Subject-specific	10,046	60.8	1.2
5	Suhaila	Comprehensive	33,475	36.5	2.1
5	Linwood	Subject-specific	28,713	48.0	2.0
6	Kimberly	Comprehensive	35,900	41.6	1.7
6	Renato	Subject-specific	30,296	44.0	3.2
7	Rilla	Comprehensive	15,029	64.6	1.0
7	Samson	Subject-specific	16,132	63.7	1.9
8	Ciara	Comprehensive	69,156	56.3	3.7
8	Placido	Subject-specific	42,075	34.9	4.7
9	Ross	Comprehensive	72,345	48.2	3.5
9	Sulayman	Subject-specific	66,234	47.1	4.3
10	Thorsten	Comprehensive	103,195	45.7	6.1
10	Adil	Subject-specific	41,801	49.6	6.3
11	Junayd	Comprehensive	61,402	62.4	5.6
11	Walker	Subject-specific	42,473	64.6	5.8
12	Everett	Comprehensive	47,773	62.0	11.2
12	Barnaby	Subject-specific	45,600	62.1	14.0
13	Sunita	Comprehensive	187,132	56.7	11.3
13	Rukiye	Subject-specific	191,934	59.4	13.6
14	Husniya	Comprehensive	208,272	60.1	12.4
14	Giselle	Subject-specific	266,663	61.5	10.8
15	Amias	Comprehensive	2,613	49.7	2.4
15	Katlyn	Subject-specific	6,061	47.7	4.6

Notes: Fictitious school district names. [a]Course Enrollment Survey 3, 2014-15 (FLDOE, 2015b). [b]Student Enrollment by District 2012-17 (FLDOE, 2017i). [c]Economic Status by District 2012-17 (FLDOE, 2017i). [d]ELL Students by District 2012-17 (FLDOE, 2017i).

Table 92

Research Questions 2 and 3 School District Paired Sample, 2016 (N = 15 school district pairs)

Pair number	School district	Course type[a]	Student population[b]	Low SES (%)[c]	EL (%)[d]
1	Judith	Comprehensive	1,697	37.6	4.2
1	Katlyn	Subject-specific	6,057	46.4	5.0
2	Antigone	Comprehensive	5,143	47.3	0.1
2	Viktor	Subject-specific	8,490	48.0	9.1
3	Phokas	Comprehensive	11,279	50.6	0.9
3	Cornell	Subject-specific	10,081	58.4	1.1
4	Evander	Comprehensive	1,290	59.9	1.1
4	Emmett	Subject-specific	1,257	59.5	7.0
5	Clifford	Comprehensive	6,821	69.7	1.1
5	Lawson	Subject-specific	8,464	61.2	2.6
6	Kimberly	Comprehensive	36,832	41.6	1.9
6	Linwood	Subject-specific	29,217	47.7	2.3
7	Suhaila	Comprehensive	33,499	41.8	1.9
7	Renato	Subject-specific	30,555	44.8	3.1
8	Rilla	Comprehensive	15,162	66.0	0.9
8	Samson	Subject-specific	15,948	64.5	2.3
9	Ross	Comprehensive	72,753	50.0	3.5
9	Sulayman	Subject-specific	67,259	48.0	4.6
10	Junayd	Comprehensive	62,764	64.5	5.8
10	Walker	Subject-specific	42,727	66.9	6.0
11	Mirela	Comprehensive	62,099	63.2	18.6
11	Barnaby	Subject-specific	46,393	62.1	13.8
12	Sunita	Comprehensive	190,121	59.0	11.8
12	Rukiye	Subject-specific	197,400	64.9	14.2
13	Husniya	Comprehensive	212,542	58.6	12.3
13	Giselle	Subject-specific	270,354	60.9	11.4
14	Ciara	Comprehensive	70,544	55.9	3.9
14	Adil	Subject-specific	42,231	49.7	6.2
15	Everett	Comprehensive	48,337	54.7	11.9
15	Placido	Subject-specific	42,459	61.0	4.4

Notes: Fictitious school district names. [a]Course Enrollment Survey 3, 2015-16 (FLDOE, 2016b). [b]Student Enrollment 2012-17 (FLDOE, 2017i). [c]Economic Status 2012-17 (FLDOE, 2017i). [d]ELL Students 2012-17 (FLDOE, 2017i).

Table 93

Research Questions 4 School District Paired Sample, 2013 (*N* = 9 school district pairs)

Pair number	School district	Course type[a, b]	Student population[c]	Low SES (%)[d]	EL (%)[e]
1	Abel	Comprehensive	18,011	56.6	6.2
1	Adil	Subject-specific	41,096	52.1	5.9
2	Kimberly	Comprehensive	35,244	36.0	1.6
2	Renato	Subject-specific	29,786	39.7	2.4
3	Farida	Comprehensive	40,670	61.1	1.2
3	Walker	Subject-specific	41,990	67.1	4.8
4	Ciara	Comprehensive	67,153	55.1	4.1
4	Placido	Subject-specific	41,495	57.3	4.1
5	Everett	Comprehensive	46,165	55.2	9.6
5	Barnaby	Subject-specific	43,789	61.2	14.2
6	Ross	Comprehensive	71,228	45.4	3.2
6	Sulayman	Subject-specific	64,463	44.8	3.6
7	Abioye	Comprehensive	85,765	65.3	7.7
7	Branch	Subject-specific	96,937	66.6	10.6
8	Sunita	Comprehensive	179,514	54.7	11.3
8	Giselle	Subject-specific	260,226	56.9	9.9
9	Husniya	Comprehensive	200,466	57.5	12.3
9	Rukiye	Subject-specific	183,066	62.1	13.5

Notes: Fictitious school district names are used. [a]Course Enrollment by School, Survey 3, 2012-13 (FLDOE, 2013a). [b]Student Enrollment by District 2012-17 (FLDOE, 2017i). [c]Economic Status by District 2012-17 (FLDOE, 2017i). [d]ELL Students by District 2012-17 (FLDOE, 2017i).

Table 94

Research Questions 4 School District Paired Sample, 2014 (N = 10 school district pairs)

Pair number	School district	Course type[a, b]	Student population[c]	Low SES (%)[d]	EL (%)[e]
1	Lavender	Comprehensive	26,850	58.0	1.8
1	Linwood	Subject-specific	28,001	52.8	1.9
2	Kimberly	Comprehensive	35,487	41.1	1.6
2	Renato	Subject-specific	30,026	41.7	2.6
3	Suhaila	Comprehensive	33,514	39.9	1.9
3	Adil	Subject-specific	41,277	44.8	6.1
4	Mirela	Comprehensive	58,175	65.4	17.9
4	Barnaby	Subject-specific	44,747	61.9	14.4
5	Ciara	Comprehensive	67,779	52.7	4.4
5	Placido	Subject-specific	41,591	57.9	4.6
6	Everett	Comprehensive	46,788	58.3	9.9
6	Branch	Subject-specific	97,972	59.8	10.8
7	Junayd	Comprehensive	60,908	61.8	5.5
7	Walker	Subject-specific	42,017	54.7	5.3
8	Ross	Comprehensive	71,024	44.7	3.3
8	Sulayman	Subject-specific	64,891	45.6	3.9
9	Sunita	Comprehensive	183,452	59.7	11.2
9	Giselle	Subject-specific	262,755	59.9	10.2
10	Husniya	Comprehensive	203,988	60.0	12.2
10	Rukiye	Subject-specific	187,274	61.1	13.1

Notes: Fictitious school district names are used. [a]Course Enrollment by School, Survey 3, 2013-14 (FLDOE, 2014b). [b]Student Enrollment by District 2012-17 (FLDOE, 2017i). [c]Economic Status by District 2012-17 (FLDOE, 2017i). [d]ELL Students by District 2012-17 (FLDOE, 2017i).

Table 95

Research Questions 4 School District Paired Sample, 2015 (*N* = 9 school district pairs)

Pair number	School district	Course type[a, b]	Student population[c]	Low SES percentage[d]	EL percentage[e]
1	Suhaila	Comprehensive	33,475	36.5	2.1
1	Linwood	Subject-specific	28,713	48.0	2.0
2	Kimberly	Comprehensive	35,900	41.6	1.7
2	Renato	Subject-specific	30,296	44.0	3.2
3	Ciara	Comprehensive	69,156	56.3	3.7
3	Placido	Subject-specific	42,075	34.9	4.7
4	Ross	Comprehensive	72,345	48.2	3.5
4	Sulayman	Subject-specific	66,234	47.1	4.3
5	Thorsten	Comprehensive	103,195	45.7	6.1
5	Adil	Subject-specific	41,801	49.6	6.3
6	Junayd	Comprehensive	61,402	62.4	5.6
6	Walker	Subject-specific	42,473	64.6	5.8
7	Everett	Comprehensive	47,773	62.0	11.2
7	Barnaby	Subject-specific	45,600	62.1	14.0
8	Sunita	Comprehensive	187,132	56.7	11.3
8	Rukiye	Subject-specific	191,934	59.4	13.6
9	Husniya	Comprehensive	208,272	60.1	12.4
9	Giselle	Subject-specific	266,663	61.5	10.8

Notes: Fictitious school district names are used. [a]Course Enrollment by School, Survey 3, 2014-15 (FLDOE, 2015b). [b]Student Enrollment by District 2012-17 (FLDOE, 2017i). [c]Economic Status by District 2012-17 (FLDOE, 2017i). [d]ELL Students by District 2012-17 (FLDOE, 2017i).

Table 96

Research Questions 4 School District Sample, Year 2016 (N = 9 school district pairs)

Pair number	School district	Course type[a, b]	Total student population[c]	Low SES percentage[d]	EL percentage[e]
1	Kimberly	Comprehensive	36,832	41.6	1.9
1	Linwood	Subject-specific	29,217	47.7	2.3
2	Suhaila	Comprehensive	33,499	41.8	1.9
2	Renato	Subject-specific	30,555	44.8	3.1
3	Ross	Comprehensive	72,753	50.0	3.5
3	Sulayman	Subject-specific	67,259	48.0	4.6
4	Junayd	Comprehensive	62,764	64.5	5.8
4	Walker	Subject-specific	42,727	66.9	6.0
5	Mirela	Comprehensive	62,099	63.2	18.6
5	Barnaby	Subject-specific	46,393	62.1	13.8
6	Sunita	Comprehensive	190,121	59.0	11.8
6	Rukiye	Subject-specific	197,400	64.9	14.2
7	Husniya	Comprehensive	212,542	58.6	12.3
7	Giselle	Subject-specific	270,354	60.9	11.4
8	Ciara	Comprehensive	70,544	55.9	3.9
8	Adil	Subject-specific	42,231	49.7	6.2
9	Everett	Comprehensive	48,337	54.7	11.9
9	Placido	Subject-specific	42,459	61.0	4.4

Notes: Fictitious school district names are used. [a]Course Enrollment by School, Survey 3, 2015-16 (FLDOE, 2016b). [b]Student Enrollment by District 2012-17 (FLDOE, 2017i). [c]Economic Status by District 2012-17 (FLDOE, 2017i). [d]ELL Students by District 2012-17 (FLDOE, 2017i).

APPENDIX E: SCHOOL-LEVEL MEAN SCALE SCORE ANALYSES

2013 Independent Samples Tests of School Mean Scale Scores for All Students

The 2013 8th grade FCAT 2.0 Science/SSA mean scale score for all students in schools that offered comprehensive science courses was numerically lower than that of schools that offered subject-specific science courses, as shown in Table 78. The 2013 FCAT 2.0 Science/SSA school mean scale score distributions for schools that offered either type of science course were sufficiently normal for the purposes of conducting a *t*-test, i.e., skew < |2.0| and kurtosis < |9.0| (Schmider et al., 2010). The assumption of equality of variances was satisfied by Levene's test, $F(571) = 1.07$, $p = .30$ (George & Mallery, 2010). The independent samples *t*-test, shown in Table 78, revealed no statistically significant difference in 2013 8th grade FCAT 2.0 Science/SSA school mean scale score for all students between groups of schools that offered comprehensive science courses and those that offered subject-specific science courses.

Table 97

2013 Independent Samples Test of School Mean Scale Scores for All Students (*N* = 573 schools)

							Independent samples test				
Descriptive statistics										95% CI	
School district group	*n* (schools)	*M*	*SD*	*SEM*	Skew	Kurtosis	*t*	*df*	*p*	*LL*	*UL*
C	401	199.30	8.53	0.43	0.01	0.28	-0.85	571	.40	-2.15	0.85
S	172	199.95	8.00	0.61	0.04	0.69					

2014 Independent Samples Tests of School Mean Scale Scores for All Students

The 2014 8th grade FCAT 2.0 Science/SSA mean scale score for all students in schools that offered comprehensive science courses was numerically lower than that of schools that offered subject-specific science courses, as shown in Table 98. The 2014 FCAT 2.0 Science/SSA school mean scale score distributions for schools that offered either type of science course were sufficiently normal for the purposes of conducting a *t*-test, i.e., skew < |2.0| and kurtosis < |9.0| (Schmider et al., 2010). The assumption of equality of variances was satisfied by Levene's test, $F(578) = 2.72$, $p = .10$ (George & Mallery, 2010). The independent samples *t*-test, shown in Table 98, revealed no statistically significant difference in 2014 8th grade FCAT 2.0 Science/SSA school mean scale score for all students between groups of schools that offered comprehensive science courses and those that offered subject-specific science courses.

Table 98

2014 Independent Samples Test of School Mean Scale Scores for All Students (*N* = 580 schools)

							Independent samples test				
Descriptive statistics										95% CI	
School district group	*n* (schools)	*M*	*SD*	*SEM*	Skew	Kurtosis	*t*	*df*	*p*	*LL*	*UL*
C	410	199.90	8.58	0.42	0.01	0.32	-0.65	578	.52	-1.99	1.00
S	170	200.39	7.71	0.59	0.52	0.66					

2015 Independent Samples Tests of School Mean Scale Scores for All Students

The 2015 8th grade FCAT 2.0 Science/SSA mean scale score for all students in schools that offered comprehensive science courses was numerically lower than that of schools that offered subject-specific science courses, as shown in Table 99. The 2015 FCAT 2.0 Science/SSA school mean scale score distributions for schools that offered either type of science course were sufficiently normal for the purposes of conducting a *t*-test, i.e., skew < |2.0| and kurtosis < |9.0| (Schmider et al., 2010). The assumption of equality of variances was satisfied by

Levene's test, $F(593) = 0.10$, $p = .75$ (George & Mallery, 2010). The independent samples *t*-test, shown in Table 99, revealed no statistically significant difference in 2015 8th grade FCAT 2.0 Science/SSA school mean scale score for all students between groups of schools that offered comprehensive science courses and those that offered subject-specific science courses.

Table 99

2015 Independent Samples Test of School Mean Scale Scores for All Students ($N = 595$ schools)

							Independent samples test				
Descriptive statistics										95% CI	
School district group	*n* (schools)	*M*	*SD*	*SEM*	Skew	Kurtosis	*t*	*df*	*p*	*LL*	*UL*
C	438	199.26	8.77	0.42	0.08	0.20	-1.55	593	.12	-2.88	0.34
S	157	200.54	8.95	0.71	-0.26	1.33					

2016 Independent Samples Tests of School Mean Scale Scores for All Students

The 2016 8th grade FCAT 2.0 Science/SSA mean scale score for all students in schools that offered comprehensive science courses was numerically lower than that of schools that offered subject-specific science courses, as shown in Table 100. The 2016 FCAT 2.0 Science/SSA school mean scale score distributions for schools that offered either type of science course were sufficiently normal for the purposes of conducting a *t*-test, i.e., skew < |2.0| and kurtosis < |9.0| (Schmider et al., 2010). The assumption of equality of variances was satisfied by Levene's test, $F(579) = 1.20$, $p = .27$ (George & Mallery, 2010). The independent samples *t*-test, shown in Table 100, revealed no statistically significant difference in 2016 8th grade FCAT 2.0

Science/SSA school mean scale score for all students between groups of schools that offered comprehensive science courses and those that offered subject-specific science courses.

Table 100

2016 Independent Samples Test of School Mean Scale Scores for All Students (N = 595 schools)

Descriptive statistics							Independent samples test				
										95% CI	
School district group	*n* (schools)	*M*	*SD*	*SEM*	Skew	Kurtosis	*t*	*df*	*p*	*LL*	*UL*
C	424	199.28	8.98	0.44	0.14	0.39	-1.34	579	.18	-2.74	0.52
S	157	200.66	8.57	0.68	-0.01	0.76					

2017 Independent Samples Tests of School Mean Scale Scores for All Students

The 2017 8th grade FCAT 2.0 Science/SSA mean scale score for all students in schools that offered comprehensive science courses was numerically lower than that of schools that offered subject-specific science courses, as shown in Table 101. The 2017 FCAT 2.0 Science/SSA school mean scale score distributions for schools that offered either type of science course were sufficiently normal for the purposes of conducting a *t*-test, i.e., skew < |2.0| and kurtosis < |9.0| (Schmider et al., 2010). The assumption of equality of variances was satisfied by Levene's test, $F(568) = 1.89$, $p = .17$ (George & Mallery, 2010). The independent samples *t*-test, shown in Table 101, revealed no statistically significant difference in 2017 8th grade FCAT 2.0 Science/SSA school mean scale score for all students between groups of schools that offered comprehensive science courses and those that offered subject-specific science courses.

Table 101

2017 Independent Samples Test of School Mean Scale Scores for All Students (N = 570 schools)

Descriptive statistics							Independent samples test				
										95% CI	
School district group	n (schools)	M	SD	SEM	Skew	Kurtosis	t	df	p	LL	UL
C	405	199.28	9.46	0.47	0.08	0.36	-1.25	568	.21	-3.70	0.82
S	165	200.24	8.73	0.68	-0.25	1.03					

APPENDIX F: 2013–2017 CHANGES IN SCHOOL DISTRICT MEAN SCALE SCORES

2013 vs. 2017 School District Mean Scale Scores for All Students, Comprehensive Group

The 8th grade FCAT 2.0 Science/SSA school district mean scale score for all students in school districts that offered comprehensive science courses was numerically higher in 2013 than in 2017, as shown in Table 102. The mean scale score distributions for all students in school districts that offered comprehensive science courses were sufficiently normal for the purposes of conducting a *t*-test, i.e., skew < |2.0| and kurtosis < |9.0| (Schmider et al., 2010). The assumption of equality of variances was satisfied by Levene's test F (96) = 0.66, p = .42 (Gardner, 2001; Morgan, 1939; Pitman, 1939). The independent samples t-test, shown in Table 102, revealed no statistically significant difference between the 2013 and 2017 8th grade FCAT 2.0 Science/SSA school district mean scale scores for all students in school districts that offered comprehensive science courses.

Table 102

2013 vs. 2017 School District Mean Scale Scores for All Students, Comprehensive Group (N = 98 school districts)

								Independent samples test				
Descriptive statistics											95% CI	
Year	*n*	*M*	*Mdn*	*SD*	*SEM*	Skew	Kurtosis	*t*	*df*	*p*	*LL*	*UL*
2013	49	198.94	199.00	5.84	0.83	-0.76	0.97	0.27	96	.79	-2.20	2.90
2017	49	198.59	199.00	6.85	0.98	-0.71	0.54					

2013 vs. 2017 School District Mean Scale Scores for All Students, Subject-specific Group

The FCAT 2.0 Science/SSA school district mean scale score for all students in school districts that offered subject-specific science courses was numerically lower in 2013 than in

2017, as shown in Table 103. The mean scale score distributions for all students in school districts that offered subject-specific science courses were sufficiently normal for the purposes of conducting a *t*-test, i.e., skew < |2.0| and kurtosis < |9.0| (Schmider et al., 2010). The assumption of equality of variances was satisfied by Levene's test F (33) = 1.18, p = .28 (Gardner, 2001; Morgan, 1939; Pitman, 1939). The independent samples t-test, shown in Table 73, revealed no statistically significant difference between the 2013 and 2017 8th grade FCAT 2.0 Science/SSA school district mean scale scores for all students in school districts that offered subject-specific science courses.

Table 103

2013 vs. 2017 School District Mean Scale Scores for All Students, Subject-specific Group (*N* = 35 school districts)

								Independent samples test				
Descriptive statistics											95% CI	
Year	*n*	*M*	*Mdn*	*SD*	*SEM*	Skew	Kurtosis	*t*	*df*	*p*	*LL*	*UL*
2013	18	199.31	199.56	5.98	1.41	-1.47	4.65	-0.92	33	.37	-4.85	1.84
2017	17	200.81	200.63	3.26	0.79	0.40	0.65					

2013 vs. 2017 School District Mean Scale Scores for Low SES Students, Comprehensive Group

The FCAT 2.0 Science/SSA school district mean scale score for low SES students in school districts that offered comprehensive science courses was numerically lower in 2013 than in 2017. As shown in Table 104, the mean scale score distributions for low SES students in school districts that offered comprehensive science courses were sufficiently normal for the purposes of conducting a *t*-test, i.e., skew < |2.0| and kurtosis < |9.0| (Schmider et al., 2010). The assumption of equality of variances was satisfied by Levene's test F (96) = 1.58, p = .21

(Gardner, 2001; Morgan, 1939; Pitman, 1939). The independent samples *t*-test, shown in Table 104, revealed no statistically significant difference between the 2013 and 2017 8th grade FCAT 2.0 Science/SSA school district mean scale score for low SES students in school districts that offered comprehensive science courses.

Table 104

2013 vs. 2017 School District Mean Scale Scores for Low SES Students, Comprehensive Group (*N* = 98 school districts)

								Independent samples test				
Descriptive statistics											95% CI	
Year	*n*	*M*	*Mdn*	*SD*	*SEM*	Skew	Kurtosis	*t*	*df*	*p*	*LL*	*UL*
2013	49	190.67	190.07	7.42	1.06	0.14	-1.09	-0.09	96	.93	-2.92	2.66
2017	49	190.78	189.06	6.44	0.92	0.41	-1.04					

2013 vs. 2017 School District Mean Scale Scores for Low SES Students, Subject-Specific Group

The FCAT 2.0 Science/SSA school district mean scale score for low SES students in school districts that offered subject-specific science courses was numerically lower in 2013 than in 2017. As shown in Table 105, the mean scale score distributions for school districts that offered subject-specific science courses met the assumption of normality for conducting a *t*-test, i.e., skew < |2.0| and kurtosis < |9.0| (Schmider et al., 2010). The assumption of equality of variances was satisfied by Levene's test, F (33) = 1.18, p = .29 (Gardner, 2001; Morgan, 1939; Pitman, 1939). The independent samples *t*-test, shown in Table 105, revealed no statistically significant difference between the 2013 and 2017 8th grade FCAT 2.0 Science/SSA school district mean scale score for low SES students in school districts that offered subject-specific science courses.

Table 105

2013 vs. 2017 School District Mean Scale Scores for Low SES Students, Subject-Specific Group (*N* = 35 school districts)

Descriptive statistics								Independent samples test				
											95% CI	
Year	*n*	*M*	*Mdn*	*SD*	*SEM*	Skew	Kurtosis	*t*	*df*	*p*	*LL*	*UL*
2013	18	188.87	189.05	4.69	1.10	0.32	-0.23	-0.71	33	.49	-4.76	2.31
2017	17	190.08	187.96	5.58	1.35	0.74	-1.10					

2013 vs. 2017 School District Mean Scale Scores for ELs, Comprehensive Group

As shown in Table 106, the FCAT 2.0 Science/SSA school district mean scale score for ELs in school districts that offered comprehensive science courses was numerically lower in 2013 than in 2017. The mean scale score distributions for ELs in school districts that offered comprehensive science courses were sufficiently normal for the purposes of conducting a *t*-test, i.e., skew < |2.0| and kurtosis < |9.0| (Schmider et al., 2010). The assumption of equality of variances was satisfied by Levene's test *F* (45) = 0.01, *p* = .94 (Gardner, 2001; Morgan, 1939; Pitman, 1939). The independent samples *t*-test, shown in Table 106, revealed that the 2017 8th grade FCAT 2.0 Science/SSA school district mean scale score for ELs in school districts that offered comprehensive science courses was statistically significantly higher than the 2013 8th grade FCAT 2.0 Science/SSA school district mean scale score for ELs in school districts that offered comprehensive science courses. The effect size was medium, based on Cohen's (1988) guidelines.

Table 106

2013 vs. 2017 School District Mean Scale Scores for ELs, Comprehensive Group (*N* = 47 school districts)

								Independent samples tests					
Descriptive statistics											95% CI		
Year	*n*	*M*	*Mdn*	*SD*	*SEM*	Skew	Kurtosis	*t*	*df*	*p*	*LL*	*UL*	*d*
2013	22	175.66	178.78	4.82	0.84	-0.89	1.40	-2.14	45	.04*	-6.20	-0.18	0.49
2017	25	178.85	178.53	5.06	1.01	0.84	0.35						

2013 vs. 2017 School District Mean Scale Scores for ELs, Subject-Specific Group

The FCAT 2.0 Science/SSA school district mean scale score for ELs in school districts that offered subject-specific science courses was numerically lower in 2013 than in 2017, as shown in Table 107. The mean scale score distributions for ELs in school districts that offered subject-specific science courses met the assumption of normality for conducting a *t*-test, i.e., skew < |2.0| and kurtosis < |9.0| (Schmider et al., 2010). The assumption of equality of variances was satisfied by Levene's test (F (20) = 0.07, p = .79) (Gardner, 2001; Morgan, 1939; Pitman, 1939). The independent samples *t*-test, shown in Table 107, revealed no statistically significant difference between the 2013 and 2017 8th grade FCAT 2.0 Science/SSA school district mean scale scores for low SES students in school districts that offered subject-specific science courses.

Table 107

2013 vs. 2017 School District Mean Scale Scores for ELs, Subject-specific Group
(*N* = 22 school districts)

Descriptive statistics							Independent samples test				
										95% CI	
Year	*n*	*M*	*SD*	*SEM*	Skew	Kurtosis	*t*	*df*	*p*	*LL*	*UL*
2013	11	177.90	3.85	1.16	-1.51	2.92	-0.37	20	.71	-4.07	2.83
2017	11	178.52	3.91	1.18	-1.05	1.20					

LIST OF REFERENCES

§ 1003.41. Next Generation Sunshine State Standards, § 1003.41 § (2016). Retrieved from http://www.leg.state.fl.us/statutes/

§1002.33 (6)(a)2, Fla. Stat. Public K-12 Education Charter Schools (2016). Retrieved from http://www.leg.state.fl.us/statutes/

§1003.02(1)(d), Fla. Stat. Public K-12 Education (2016). Retrieved from http://www.leg.state.fl.us/statutes/

§1008.31 Fla. Stat. (2016). Florida's K-20 education performance accountability system; legislative intent; mission, goals, and systemwide measures; data quality improvements. Retrieved September 9, 2017, from http://www.leg.state.fl.us/statutes/

§1008.34 Fla. Stat. Assessment and accountability (2016). Retrieved from http://www.leg.state.fl.us/statutes/

Abedi, J. (2002). Standardized achievement tests and English language learners: Psychometrics issues. *Educational Assessment*, *8*(3), 231–257. https://doi.org/10.1207/S15326977EA0803_02

Abedi, J. (2003). *Impact of student language background on content-based performance: Analyses of extant data.* (No. CSE-R-603). Los Angeles: National Center for Research on Evaluation, Standard and Student Testing, Center for the Study of Evaluation, UCLA. Retrieved from https://eric.ed.gov/?id=ED480903

Abrami, P. C., Poulsen, C., & Chambers, B. (2004). Teacher motivation to implement an educational innovation: Factors differentiating users and non-users of cooperative learning. *Educational Psychology*, *24*(2), 201–216.

ACT, Inc. (2016). *The condition of STEM 2016* (No. MS501) (p. 31). Iowa City, IA: ACT, Inc. Retrieved from http://www.act.org/content/act/en/research/condition-of-stem-2016.html

Adler, S., Dougan, A., & Garcia, J. (2006). NCATE has a lot to say to future social studies teachers: A response to Sam Wineburg. *The Phi Delta Kappan*, *87*(5), 396–400.

Alwardt, R. K. (2011). *Investigating the transition process when moving from a spiral curriculum alignment into a field-focus science curriculum alignment in middle school* (Dissertation). Linwood University, St. Charles, MO.

Amah, E., Daminabo-Weje, M., & Dosunmu, R. (2013). Size and organizational effectiveness: Maintaining a balance. *Advances in Management and Applied Economics*, *3*(5), 115.

Amaral, O. M., Garrison, L., & Klentschy, M. (2002). Helping English learners increase achievement through inquiry-based science instruction. *Bilingual Research Journal*, *26*(2), 213.

American Association for the Advancement of Science. (1989). *Science for all Americans: A Project 2061 report on literacy goals in science, mathematics, and technology*. Washington, DC: AAAS Publication, no. 89-01S.

American Association for the Advancement of Science. (1993). *Benchmarks for science literacy*. New York, NY: Oxford University Press.

American Association for the Advancement of Science. (2000). *Designs for science literacy*. New York, NY: Oxford University Press.

American Association for the Advancement of Science. (2001). *Atlas of science literacy*. Washington, DC: National Science Teachers Association.

American Association for the Advancement of Science. (2009). *Benchmarks online: Project 2061*. Washington, DC: American Association for the Advancement of Science. Retrieved from http://www.project2061.org/publications/bsl/online/index.php

American Association for the Advancement of Science. (2011). Project 2061. Retrieved October 15, 2017, from http://www.project2061.org/research/assessment.htm

Anderson, L., & Krathwohl, D. R. (Eds.). (2001). *A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives* (Vol. 41). New York, NY: Addison Wesley Longman, Inc. Retrieved from http://www.tandfonline.com/doi/pdf/10.1207/s15430421tip4104_2

Anderson, R. D., & Helms, J. V. (2001). The ideal of standards and the reality of schools: Needed research. *Journal of Research in Science Teaching*, *38*(1), 3.

Anelli, C. (2011). Scientific literacy: What is it, are we teaching it, and does it matter. *American Entomologist*, *57*(4), 235–244.

Anft, M. (2013). The STEM crisis: Reality or myth. *The Chronicle of Higher Education*, *58*(12), 1–14.

Arens, A. K., Pekrun, R., Murayama, K., Marsh, H. W., Lichtenfeld, S., & Hofe, R. (2017). Math self-concept, grades, and achievement test scores: Long-term reciprocal effects across five waves and three achievement tracks. *Journal of Educational Psychology*, *109*(5), 621–634. https://doi.org/10.1037/edu0000163.supp

Aschbacher, P., Ing, M., & Tsai, S. (2014). Is science me? Exploring middle school students' STEM career aspirations. *Journal of Science Education & Technology*, *23*(6), 735–743. https://doi.org/10.1007/s10956-014-9504-x

Åström, M. (2007). *Integrated and subject-specific: An empirical exploration of science education in Swedish compulsory schools*. Swedish National Graduate School in Science and Technology Education, FontD: Department of Social and Welfare Studies, Linköping University, Norrköping.

Åström, M. (2008). Defining integrated science education and putting it to test. Retrieved from http://www.diva-portal.org/smash/record.jsf?pid=diva2:422913

Åström, M., & Karlsson, K.-G. (2012). Using hierarchical linear models to test differences in Swedish results from OECD's PISA 2003: Integrated and subject-specific science education. *Nordic Studies in Science Education*, *3*(2), 121–131.

Atkin, J. M., & Black, P. (2003). *Inside science education reform: A history of curricular and policy change*. New York, NY: Teachers College Press.

Atkin, J. M., & Black, P. (2010). History of science curriculum reform in the United States and the United Kingdom. In S. K. Abell & N. G. Lederman (Eds.), *Handbook of research on science education* (pp. 781–806). New York, NY: Routledge.

Au, W. (2007). High-stakes testing and curricular control: A qualitative metasynthesis. *Educational Researcher*, *36*(5), 258–267. https://doi-org.ezproxy.net.ucf.edu/10.3102/0013189X07306523

Bandura, A. (1997). *Self-efficacy: The exercise of control*. New York, NY: W.H. Freeman.

Bandura, A. (2000). Self-efficacy: The foundation of agency. In W. J. Perrig, A. Grob, W. J. Perrig, & A. Grob (Eds.), *Control of human behavior, mental processes, and consciousness: Essays in honor of the 60th birthday of August Flammer.* (pp. 17–33). Mahwah, NJ: Lawrence Erlbaum Associates Publishers.

Banilower, E. R., Smith, P. S., Weiss, I. R., Malzahn, K. A., Campbell, K. M., Weis, A. M., & Horizon Research, Inc. (2013). *Report of the 2012 national survey of science and mathematics education*. Horizon Research, Inc.

Becker, B. E., & Luthar, S. S. (2002). Social–emotional factors affecting achievement outcomes among disadvantaged students: Closing the achievement gap. *Educational Psychologist*, *37*(4), 197–214. https://doi.org/10.1207/S15326985EP3704_1

Behind the Name. (2017). Random name generator. Retrieved October 31, 2017, from https://www.behindthename.com/random/random.php?number=1&gender=both&surnam

e=&showextra=&nodiminutives=yes&all=no&usage_afr=1&usage_alb=1&usage_ara=1&usage_eng=1&usage_ger=1&usage_ame=1

Best, J. W., & Kahn, J. V. (1998). *Research in education*. Needham Heights, MA: Allyn & Bacon. Retrieved from http://ww2.odu.edu/~jritz/attachments/reined.pdf

Betz, N. E., Hackett, G., Lent, R. W., Lopez, F. G., & Bieschke, K. J. (1991). Mathematics self-efficacy-college courses scales. *Mathematics Self-Efficacy: Sources and Relation to Science-Based Career Choice*, *38*, 424–430.

Bing Wei. (2009). In search of meaningful integration: The experiences of developing integrated science curricula in junior secondary schools in China. *International Journal of Science Education*, *31*(2), 259–277. https://doi.org/10.1080/09500690701687430

Bjork, R. A. (1994). Metacognition: Knowing about knowing. In J. Metcalfe & A. Simamura (Eds.), *Memory and metamemory considerations in the training of human beings* (pp. 185–205). Cambridge, MA: MIT Press.

Black, P., & Wiliam, D. (2004). Classroom assessment is not (necessarily) formative assessment (and vice-versa). *Yearbook of the National Society for the Study of Education*, *103*(2), 183.

Bohrnstedt, G. (2016). TIMSS, PISA, and NAEP: What to know before digging into the results [Text]. Retrieved September 3, 2017, from http://www.air.org/resource/timss-pisa-and-naep-what-know-digging-results

Borman, G. D., & Kimball, S. M. (2005). Teacher quality and educational equality: Do teachers with higher standards-based evaluation ratings close student achievement gaps? *The Elementary School Journal*, *106*(1), 3–20. https://doi.org/10.1086/496904

Boser, U. (2013). *Size matters: A look at school district consolidation*. Center for American Progress. Retrieved from https://www.americanprogress.org/wp-content/uploads/2013/08/SchoolDistrictSize.pdf

Bouck, E. C. (2004). How size and setting impact education in rural schools. *Rural Educator*, *25*(3), 38–42.

Bower, G. H., & Hilgard, E. R. (1981). *Theories of learning*. Englewood Cliffs, NJ: Prentice-Hall. Retrieved from http://garfield.library.upenn.edu/classics1988/A1988L489700001.pdf

Bower, J. M. (2005). *Scientists and science education reform: Myths, methods, and madness*. Washington, DC: National Academy of Sciences. Retrieved from http://www.nas.edu/rise/backg2a.htm

Branch, G. (2008). Science, evolution, and creationism. *Reports of the National Center for Science Education*, *28*(1), 14.

Bransford, J. D., Brown, A. L., & Cocking, R. R. (Eds.). (2000). *How people learn* (expanded). Washington, DC: National Academy Press.

Brayboy, B. M. J., Faircloth, S. C., Lee, T. S., Maaka, M. J., & Richardson, T. A. (2015). Sovereignty and education: An overview of the unique nature of indigenous education. *Journal of American Indian Education*, *54*(1), 1–9.

Britner, S. L., & Pajares, F. (2006). Sources of science self-efficacy beliefs of middle school students. *Journal of Research in Science Teaching*, *43*(5), 485–499. https://doi.org/10.1002/tea.20131

Britton, E. D., & Schneider, S. A. (2010). Large-scale assessments in science education. In S. K. Abell & N. G. Lederman (Eds.), *Handbook of research on science education* (pp. 1007–1040). New York, NY: Routledge.

Brooks, J. G., & Brooks, M. C. (1999). *In search of understanding: The case for constructivist classrooms* (Revised). Alexandria, VA: Association for Supervision and Curriculum Development.

Brown, B. A., Reveles, J. M., & Kelly, G. J. (2005). Scientific literacy and discursive identity: A theoretical framework for understanding science learning. *Science Education*, *89*(5), 779–802. https://doi.org/10.1002/sce.20069

Brown, T. A. (2006). *Confirmatory factor analysis for applied research*. New York, NY: Guilford Press.

Bybee, R. W. (1995). Science curriculum reform in the United States. In R. W. Bybee & J. D. McInerney (Eds.), *Redesigning the science curriculum*. Colorado Springs, CO: Biological Sciences Curriculum Study. Retrieved from http://www.nas.edu/rise/backg3a.htm

Bybee, R. W. (1997). *Achieving scientific literacy*. Portsmouth, NH: Heinemann Educational Books.

Carnoy, M. (2015). *International test score comparisons and educational policy: A review of the critiques*. Boulder, CO: National Education Policy Center. Retrieved from http://nepc.colorado.edu/files/pb_carnoy_international_test_scores_0.pdf

Carnoy, M., & Rothstein, R. (2013). *What do international tests really show about U.S. student performance?* (p. 99). Washington, DC: Economic Policy Institute. Retrieved from http://www.epi.org/publication/us-student-performance-testing/

Carr, J. F., & Harris, D. E. (2001). *Succeeding with standards: Linking curriculum, assessment, and action planning* (eBook Collection EBSCOhost). Alexandria, VA: Assoc. for Supervision and Curriculum Development. Retrieved from http://eds.b.ebscohost.com.ezproxy.net.ucf.edu/eds/ebookviewer/ebook/bmxlYmtfXzc0MTUxX19BTg2?sid=8ea37c42-8371-461c-a3b4-e8241a81f99f@sessionmgr103&vid=13&format=EB&rid=1

Champagne, A. B. (1997). Science education in the United States. *Frontiers: The Interdisciplinary Journal of Study Abroad*, *3*(2), 52–63.

Charette, R. N. (2013). The STEM crisis is a myth. *Institute of Electrical and Electronics EEE Spectrum*, *50*(9), 44. https://doi.org/10.1109/MSPEC.2013.6587189

Chen, X., & Soldner, M. (2013). *STEM attrition: College students' paths into and out of STEM fields* (p. 49). Washington, DC: National Center for Education Statistics, Institute of Education Sciences, U.S. Dept. of Education. Retrieved from https://nces.ed.gov/pubs2014/2014001rev.pdf

Cirio, J. (2017). Request for FCAT 2.0 science/statewide science assessment technical reports.

Clifford, B. A. (2016). *Middle school science curriculum design and 8th grade student achievement in Massachusetts public schools* (Dissertation). Capella, Minneapolis, MN.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Collier, L. (2017). *What does ESSA mean for assessment?* Princeton, NJ: Educational Testing Service. Retrieved from http://news.ets.org/stories/what-does-essa-mean-for-assessment/

Cosentino de Cohen, C., Deterding, N., & Clewell, B. C. (2005). *Who's left behind? Immigrant children in high and low LEP schools.* Washington, DC: The Urban Institute, Program for Evaluation and Equity Research. Retrieved from https://eric.ed.gov/?id=ED490928

Czerniak, C. M. (2007). Interdisciplinary science teaching. In S. K. Abell & N. G. Lederman (Eds.), *Handbook of research on science education* (pp. 537–559). Mahwah, N.J: Lawrence Erlbaum Associates.

David, J. L. (2011). What students need to learn: High-stakes testing narrows the curriculum. *Educational Leadership*, *68*(6), 78–80.

Davis, K. S. (2003). "Change is hard": What science teachers are telling us about reform and teacher learning of innovative practices. *Science Education*, *87*(1), 3–30. https://doi.org/10.1002/sce.10037 de Ayala, R. J. (2008). *The theory and practice of item response theory*. New York, NY: The Guildford Press. Retrieved from https://tech.knewton.com/blog/2012/06/understanding-student-performance-with-item-response-theory/

De Winter, J. C. (2013). Using the Student's t-test with extremely small sample sizes. *Practical Assessment, Research & Evaluation*, *18*(10).

DeBard, R., & Kubow, P. (2002). From compliance to commitment: The need for constituent discourse in implementing testing policy. *Educational Policy*, *16*(3), 387–405.

DeBoer, G. E. (1991). *A history of ideas in science education: Implications for practice*. New York, NY: Teachers College Press.

DeBoer, G. E. (2000). Scientific literacy: Another look at its historical and contemporary meanings and its relationship to science education reform. *Journal of Research in Science Teaching*, *37*(6), 582–601.

DeBoer, G. E. (2002). Student-centered teaching in a standards-based world: Finding a sensible balance. *Science & Education*, *11*(4), 405–417.

Desilver, D. (2017). *U.S. students' academic achievement still lags that of their peers in many other countries*. Pew Research Center. Retrieved from http://www.pewresearch.org/fact-tank/2017/02/15/u-s-students-internationally-math-science/

Dewey, J. (1902). *The child and the curriculum*. Chicago, IL: University of Chicago Press.

Dewey, J. (1910). *How we think*. Boston: D. C. Heath & Co.

Dewey, J. (1916). Method in science teaching. *Science Education*, *1*(1), 3.

Dewey, J. (1938). *Experience and education*. New York, NY: Macmillan.

Diehl, D. E. (2005). *A study of faculty-related variables and competence in integrating instructional technologies into pedagogical practices*. Retrieved from http://debdavis.pbworks.com/w/file/fetch/94210478/Curriculum_and_Biblical_Perspectives_Chart_progress.docx

Driscoll, D., Halcoussis, D., & Svorny, S. (2003). School district size and student performance. *Economics of Education Review*, *22*(2), 193–201. https://doi.org/10.1016/S0272-7757(02)00002-X

Drits-Esser, D., & Stark, L. A. (2015). The impact of collaborative curriculum design on teacher professional learning. *Electronic Journal of Science Education*, *19*(8), 1–27.

DuFour, R. (2011). Professional learning communities: A bandwagon, an idea worth considering or our best hope for high levels of learning? *Counterpoints*, *408*, 159–164.

Duncan, A. (2010). Secretary Arne Duncan's remarks at OECD's release of the Program for International Student Assessment (PISA) 2009 Results. U.S. Department of Education. Retrieved from http://www.ed.gov/news/speeches/secretary-arne-duncans-remarks-oecds-release-program-international-student-assessment-

Engelhart, M. D., Furst, E. J., Hill, W. H., & Krathwohl, D. R. (1956). *Taxonomy of educational objectives, handbook I: The cognitive domain*. (B. S. Bloom, Ed.). New York, NY: David McKay Co., Inc.

Every Student Succeeds Act, Pub. L. No. 114–95, § 1177 (2015). Retrieved from https://www.congress.gov/bill/114th-congress/senate-bill/1177/text

Fathman, A. K., Kessler, C., & Quinn, M. E. (1992). *Teaching science to English learners, grades 4-8*. Washington, DC: National Clearinghouse for Bilingual Education.

Faulkner, S. F. (2012). Science literacy: Exploring middle-level science curriculum structure and student achievement. In *NERA Conference Proceedings* (Vol. 18). ProQuest LLC.

Fla. Admin. Code R. 6A-1.09422. Statewide Standardized Assessment Program Requirements (2016). Retrieved from https://www.flrules.org/gateway/RuleNo.asp?id=6A-1.09422

Fla. Admin. Code R. 6A-1.09981. School and District Accountability (2016). Retrieved from https://www.flrules.org/gateway/ruleNo.asp?ID=6A-1.09981

Fleischman, H. L., Hopstock, P. J., Pelczar, M. P., & Shelley, B. E. (2010). *Highlights from PISA 2009: Performance of U.S. 15-year-old students in reading, mathematics, and science literacy in an international context*. Washington, DC: National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education. Retrieved from http://eric.ed.gov/?id=ED513640

Florida Consortium of Public Charter Schools. (2017). What is a charter school? Retrieved March 16, 2017, from http://floridacharterschools.org/schools/what_is_a_charter_school/

Florida Department of Education. (2008a). 2008 science standards overview. Florida Department of Education. Retrieved from http://www.cpalms.org/Uploads/docs/FrontMatter/2008_Science%20Standards_Overview.pdf

Florida Department of Education. (2008b). *Sunshine State Standards: Science* (p. 73). Tallahassee, FL: Florida Department of Education. Retrieved from http://www.heartlanded.org/mathscience/documents/science_k-12_standards.pdf

Florida Department of Education. (2012a). *English Language Learners (ELL) database and program handbook*. Retrieved from http://www.fldoe.org/core/fileparse.php/7574/urlt/0101166-edph1112.pdf

Florida Department of Education. (2012b). FCAT 2.0 parent information sheet, developmental scale scores. Retrieved March 16, 2017, from http://www.fldoe.org/core/fileparse.php/3/urlt/pidss-final.pdf

Florida Department of Education. (2012c). *FCAT 2.0 Science test item specifications version 2 grade 8*. Tallahassee, FL: Florida Department of Education. Retrieved from http://www.fldoe.org/core/fileparse.php/5682/urlt/0077914-fl09g8sci.pdf

Florida Department of Education. (2012d). *Science standards*. Tallahassee, FL. Retrieved from http://www.cpalms.org/Downloads.aspx

Florida Department of Education. (2013a). Course Enrollment, Survey 3, 2012-13. Retrieved March 16, 2017, from http://www.fldoe.org/accountability/assessments/k-12-student-assessment/history-of-fls-statewide-assessment/archived-publications.stml

Florida Department of Education. (2013b). *FCAT 2.0 Florida statewide assessments 2013 yearbook* (p. 668). Tallahassee, FL: Florida Department of Education.

Florida Department of Education. (2013c). NGSSS test design summary operational assessments. Retrieved March 16, 2017, from http://www.fldoe.org/core/fileparse.php/5662/urlt/NGSSS_TestDesignSummary_Final.pdf

Florida Department of Education. (2013d). *Understanding FCAT 2.0 Reports, Spring 2013* (p. 31). Tallahassee, FL: Florida Department of Education. Retrieved from http://www.fldoe.org/core/fileparse.php/3/urlt/2013ufr.pdf

Florida Department of Education. (2014a). Charter Schools. Retrieved July 7, 2017, from http://www.fldoe.org/schools/school-choice/charter-schools/

Florida Department of Education. (2014b). Course Enrollment, Survey 3, 2013-14. Retrieved March 16, 2017, from http://www.fldoe.org/accountability/assessments/k-12-student-assessment/history-of-fls-statewide-assessment/archived-publications.stml

Florida Department of Education. (2014c). *FCAT 2.0 Florida statewide assessments 2014 technical report* (p. 175). Tallahassee, FL: Florida Department of Education. Retrieved from https://accountabaloney.files.wordpress.com/2016/05/fcat_2-0_fl_eoc_2014_technical_report_final_distribute.pdf

Florida Department of Education. (2014d). *FCAT 2.0 Florida statewide assessments 2014 yearbook* (p. 752). Tallahassee, FL: Florida Department of Education.

Florida Department of Education. (2014e). NGSSS test design summary operational assessments. Retrieved March 16, 2017, from http://www.fldoe.org/core/fileparse.php/5662/urlt/NGSSS_TestDesignSummary_Final.pdf

Florida Department of Education. (2014f). Our Mission. Retrieved July 4, 2017, from http://www.fldoe.org/schools/higher-ed/fl-college-system/data-reports/fl-colleges-accountability-reports/our-mission.stml

Florida Department of Education. (2014g). *Test items specifications*. Tallahassee, FL: Florida Department of Education. Retrieved from http://www.fldoe.org/accountability/assessments/k-12-student-assessment/history-of-fls-statewide-assessment/fcat-2-0/test-items-specifications.stml

Florida Department of Education. (2014h). *Understanding FCAT 2.0 Reports, Spring 2014*. Tallahassee, FL: Florida Department of Education. Retrieved from http://www.fldoe.org/core/fileparse.php/3/urlt/2014ufr.pdf

Florida Department of Education. (2015a). About the Florida Standards. Retrieved July 1, 2017, from http://www.fldoe.org/academics/standards/florida-standards/about.stml

Florida Department of Education. (2015b). Course Enrollment, Survey 3, 2014-15. Retrieved April 8, 2017, from http://www.fldoe.org/accountability/data-sys/edu-info-accountability-services/pk-12-public-school-data-pubs-reports/archive.stml

Florida Department of Education. (2015c). Education statistics definitions, FLDOE information database requirements Volume 1: Automated student information system, automated student data elements, 2014-15, data element number 131037. Florida Department of Education. Retrieved from https://edstats.fldoe.org/SASPortal/navigate.do?PortalPage=PortalPage%2Bomi%3A%2F%2FMETASERVER.Foundation%2Freposname%3DFoundation%2FPSPortalPage%3Bid%3DA5YWB4SY.BN00000A

Florida Department of Education. (2015d). *FCAT 2.0 Florida statewide assessments 2015 yearbook* (p. 330). Tallahassee, FL: Florida Department of Education.

Florida Department of Education. (2015e). Grades 5 and 8 Statewide Science Assessment, statewide comparison report 2012-2015. Retrieved March 16, 2017, from http://www.fldoe.org/core/fileparse.php/5668/urlt/2015ScienceStatewideComparisonRpt.pdf

Florida Department of Education. (2015f). *Understanding FCAT 2.0 Reports, Science and Reading Retake, Spring 2015*. Tallahassee, FL: Florida Department of Education. Retrieved from http://www.fldoe.org/core/fileparse.php/5663/urlt/2015UFR.pdf

Florida Department of Education. (2016a). *2015-16 guide to calculating school and district grades*. Tallahassee, FL: Florida Department of Education. Retrieved from http://schoolgrades.fldoe.org/pdf/1516/SchoolGradesCalcGuide16.pdf

Florida Department of Education. (2016b). Course Enrollment, Survey 3, 2015-16. Retrieved April 10, 2017, from http://www.fldoe.org/accountability/data-sys/edu-info-accountability-services/pk-12-public-school-data-pubs-reports/archive.stml

Florida Department of Education. (2016c). Florida Standards Assessments, Statewide Science Assessments. Retrieved April 1, 2017, from http://www.fldoe.org/accountability/assessments/k-12-student-assessment/fsa.stml

Florida Department of Education. (2016d). *Florida statewide science and EOC assessments 2016 technical report*. Tallahassee, FL: Florida Department of Education. Retrieved from fsassessments.org/wp-content/uploads/2016/04/V4_FSA_TechRpt_508.pdf

Florida Department of Education. (2016e). *Florida Statewide Science Assessment content focus report, 2016* (p. 3). Tallahassee, FL. Retrieved from http://www.fldoe.org/core/fileparse.php/5663/urlt/Science58ScienceCF16.pdf

Florida Department of Education. (2016f). NGSSS test design summary operational assessments. Retrieved March 16, 2017, from http://www.fldoe.org/core/fileparse.php/5662/urlt/NGSSS_TestDesignSummary_Final.pdf

Florida Department of Education. (2016g). *Statewide Science Assessment grades 5 & 8 science administration manual*. Tallahassee, FL: Pearson Vue, Inc.

Florida Department of Education. (2016h). Understanding NGSSS Reports Grades 5 & 8 Science and End-of-Course Assessments Spring 2016. Retrieved March 16, 2017, from http://www.fldoe.org/accountability/accountability-reporting/publications-guides/

Florida Department of Education. (2016i). *Understanding NGSSS Reports, Grades 5 & 8 Science and End-of-Course Assessments, Spring 2016* (p. 32). Tallahassee, FL: Florida Department of Education. Retrieved from http://www.fldoe.org/core/fileparse.php/5662/urlt/UNGSSSEOCRSpring16.pdf

Florida Department of Education. (2017a). 2016-17 guide to calculating school and district grades. Bureau of Accountability Reporting; Division of Accountability, Research, and Measurement. Retrieved from http://schoolgrades.fldoe.org/pdf/1617/SchoolGradesCalcGuide17.pdf

Florida Department of Education. (2017b). *2016-17 guide to calculating school and district grades* (p. 32). Tallahassee, FL: Florida Department of Education. Retrieved from http://www.fldoe.org/core/fileparse.php/18534/urlt/SchoolGradesCalcGuide17.pdf

Florida Department of Education. (2017c). 2016-17 Next Generation Sunshine State Standards Statewide Science Assessment and FCAT 2.0 Reading retake fact sheet. Florida Department of Education. Retrieved from http://www.fldoe.org/core/fileparse.php/5663/urlt/1617NGSSSScienceReadingRFS.pdf

Florida Department of Education. (2017d). 2016-2017 Course Code Directory. Retrieved March 16, 2017, from http://www.fldoe.org/policy/articulation/ccd/2016-2017-course-directory.stml

Florida Department of Education. (2017e). 2017 Statewide Science and Social Studies Assessments. Retrieved July 9, 2017, from http://www.fldoe.org/core/fileparse.php/5668/urlt/90NGSSSPacket.pdf

Florida Department of Education. (2017f). Course Enrollment, Survey 3, 2016-17. Retrieved March 16, 2017, from http://www.fldoe.org/accountability/assessments/k-12-student-assessment/history-of-fls-statewide-assessment/archived-publications.stml

Florida Department of Education. (2017g). *English Language Learners by school, by race, preliminary survey 2, 2016-17*.

Florida Department of Education. (2017h). *Florida Statewide Science Assessment content focus report, 2017* (p. 3). Tallahassee, FL: Florida Department of Education. Retrieved from http://www.fldoe.org/core/fileparse.php/5663/urlt/ScienceCF17.pdf

Florida Department of Education. (2017i). Florida's PK-12 education information portal. Retrieved April 9, 2017, from https://edstats.fldoe.org/SASWebReportStudio/gotoReportSection.do?sectionNumber=7

Florida Department of Education. (2017j). History of Florida's Statewide Assessment Program. Retrieved August 2, 2017, from http://www.fldoe.org/accountability/assessments/k-12-student-assessment/history-of-fls-statewide-assessment/history-fl-statewide-assessment.stml

Florida Department of Education. (2017k). NGSSS state report of district results, grade 8. Retrieved March 16, 2017, from http://www.fldoe.org/accountability/assessments/k-12-student-assessment/results/

Florida Department of Education. (2017l). NGSSS test design summary operational assessments. Retrieved March 16, 2017, from http://www.fldoe.org/core/fileparse.php/5662/urlt/NGSSS_TestDesignSummary_Final.pdf

Florida Department of Education. (2017m). *School grades accountability update*. Tallahassee, FL: Florida Department of Education. Retrieved from http://www.fldoe.org/core/fileparse.php/7506/urlt/Accountability-Update.pdf

Florida Department of Education. (2017n). Science standards. Retrieved March 16, 2017, from http://www.fldoe.org/academics/standards/subject-areas/math-science/science/

Florida Department of Education. (2017o). Statewide Science Assessment. Retrieved July 12, 2017, from http://www.fldoe.org/accountability/assessments/k-12-student-assessment/science.stml

Florida Department of Education. (2017p). Statewide Science Assessment Results. Retrieved April 2, 2017, from http://www.fldoe.org/accountability/assessments/k-12-student-assessment/results/2016.stml

Florida State University. (2017). CPALMS: Florida's official source for standards information and course descriptions. Retrieved March 16, 2017, from http://www.cpalms.org/Public/search/Course

Fraenkel, J. R., Wallen, N. E., & Hyun, H. (2015). *How to design and evaluate research in education* (9th ed.). Boston: McGraw-Hill Higher Education.

Friend, H. (1985). The effect of science and mathematics integration on selected seventh grade students' attitudes toward and achievement in science. *School Science & Mathematics*, *85*, 453–461. https://doi.org/10.1111/j.1949-8594.1985.tb09648.x

Fulton, K. P. (2012). 10 reasons to flip: A southern Minnesota school district flipped its math classrooms and raised achievement and student engagement. *Phi Delta Kappan*, *94*(2), 20.

Gao, S. (2014). Relationship between science teaching practices and students' achievement in Singapore, Chinese Taipei, and the US: An analysis using TIMSS 2011 data. *Frontiers of Education in China*, *9*(4), 519–551.

Gao, S., & Wang, J. (2014). Teaching transformation under centralized curriculum and teacher learning community: Two Chinese chemistry teachers' experiences in developing inquiry-based instruction. *Teaching and Teacher Education*, *44*, 1–11. https://doi.org/10.1016/j.tate.2014.07.008

Gao, S., & Wang, J. (2016). Do variations of science teaching approaches make difference in shaping student content and problem solving achievement across different racial/ethnic groups? *International Journal of Environmental and Science Education*, *11*(12), 5404–5428.

Gardner, R. C. (2001). *Psychological statistics using SPSS for Windows*. New Jersey: Prentice Hall.

Geier, R., Blumenfeld, P. C., Marx, R. W., Krajcik, J. S., Fishman, B., Soloway, 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. https://doi.org/10.1002/tea.20248

George, D., & Mallery, M. (2010). *SPSS for windows step by step: A simple guide and reference* (10a ed.). Boston, MA: Pearson.

Gibbs, W. W., & Fox, D. (1999). The false crisis in science education. *Scientific American*, *281*(4), 86–93. https://doi.org/10.2307/26058443

Goldberg, H., & Wagreich, P. (1990). A model integrated mathematics science program for the elementary school. *International Journal of Educational Research*, *14*(2), 193–214.

González-Howard, M., & McNeill, K. L. (2016). Learning in a community of practice: Factors impacting English-learning students' engagement in scientific argumentation. *Journal of Research in Science Teaching*, *53*(4), 527–553.

Gottfried, M. A., & Williams, D. N. (2013). STEM club participation and STEM schooling outcomes. *Education Policy Analysis Archives*, *21*(79). Retrieved from https://login.ezproxy.net.ucf.edu/login?auth=shibb&url=http://search.ebscohost.com/login.aspx?direct=true&db=eric&AN=EJ1018827&site=eds-live&scope=site

Guo, C.-J. (2010). Issues in science learning: An international perspective. In S. K. Abell & N. G. Lederman (Eds.), *Handbook of research on science education* (pp. 227–256). New York, NY: Routledge.

Haag, S., & Megowan, C. (2015). Next Generation Science Standards: A national mixed-methods study on teacher readiness. *School Science and Mathematics*, *115*(8), 416–426.

Hackett, G., & Betz, N. E. (1981). A self-efficacy approach to the career development of women. *Journal of Vocational Behavior*, *18*(3), 326–339. https://doi.org/10.1016/0001-8791(81)90019-1

Halpern, M., & Halpern, D. F. (2005). Making transfer happen across physical, temporal, and conceptual space. In J. Mestre (Ed.), *Transfer of learning: From a modern multidisciplinary perspective* (pp. 357–370). Greenwich, CT: Information Age Publishing.

Han, S. W. (2016). Standards-based external exams and students' science-related career expectations: An international perspective. *Educational Research & Evaluation*, *22*(7/8), 374–401. https://doi.org/10.1080/13803611.2016.1257946

Hannaway, J., & Kimball, K. (1998). Big isn't always bad: School district size, poverty, and standards-based reform. Retrieved from https://eric.ed.gov/?id=ED439184

Hanushek, E. A. (2010). How well do we understand achievement gaps? *Focus*, *27*(2), 5–12.

Hanushek, E. A., Peterson, P. E., & Woessmann, L. (2014). U.S. students from educated families lag in international tests. *Education Next*, *14*(4), 8–18.

Hargrove, T. Y., Jones, M. G., Jones, B. D., Hardin, B., Chapman, L., & Davis, M. (2000). Unintended consequences of high-stakes testing in North Carolina: Teacher perceptions. *Educational Research Service Spectrum*, *18*(4), 21–25.

Harlen, W., & Deakin Crick, R. (2003). Testing and motivation for learning. *Assessment in Education: Principles, Policy & Practice*, *10*(2), 169–207. https://doi.org/10.1080/0969594032000121270

Hart, A. (2001). Mann-Whitney test is not just a test of medians: Differences in spread can be important. *British Medical Journal*, *323*(7309), 391–393.

Hartzler, D. S. (2000). *A meta-analysis of studies conducted on integrated curriculum programs and their effects on student achievement* (Dissertation). Indiana University, Bloomington, IN. Retrieved from http://ir.library.oregonstate.edu/xmlui/handle/1957/55648

Hattie, J. (1985). Methodology review: Assessing unidimensionality of tests and items. *Applied Psychological Measurement*, *9*(2), 139–164.

Hattie, J. (2009). *Visible learning: A synthesis of over 800 meta-analyses relating to achievement*. New York, NY: Routledge.

Hattie, J. (2017). Visible learning plus. Retrieved March 17, 2017, from http://visiblelearningplus.com/content/faq

Haught, B. (2008). Evolution in the Sunshine State: The fight over evolution in the state science standards. *Reports of the National Center for Science Education*, *28*(4), 28–31.

Herr, N. (2007). Layered & spiral curriculum. In *The sourcebook for teaching science* (p. 584). San Francisco, CA: John Wiley/Jossey-Bass Publishers. Retrieved from http://www.csun.edu/science/books/sourcebook/chapters/24-curriculum/graphics/layered-spiraling.html

Herrington, D., & Daubenmire, P. L. (2016). No teacher is an island: Bridging the gap between teachers' professional practice and research findings. *Journal of Chemical Education*, *93*(8), 1371–1376.

Hess, K. K., Jones, B. S., Carlock, D., & Walkup, J. R. (2009). Cognitive rigor: Blending the strengths of Bloom's taxonomy and Webb's depth of knowledge to enhance classroom-level processes. *Online Submission*. Retrieved from https://eric.ed.gov/?id=ED517804

Ho, A. D. (2007). Discrepancies between score trends from NAEP and state tests: A scale-invariant perspective. *Educational Measurement: Issues & Practice*, *26*(4), 11–20. https://doi.org/10.1111/j.1745-3992.2007.00104.x

Hoeg, D. G., & Bencze, J. L. (2017). Values underpinning STEM education in the USA: An analysis of the Next Generation Science Standards. *Science Education*, *101*(2), 278–301. https://doi.org/10.1002/sce.21260

Holbrook, J., & Rannikmae, M. (2009). The meaning of scientific literacy. *International Journal of Environmental and Science Education*, *4*(3), 275–288.

Hollweg, K. S., & Hill, D. (2003). *What is the influence of the National Science Education Standards? Reviewing the evidence*. Washington, DC: National Academies Press. Retrieved from https://login.ezproxy.net.ucf.edu/login?auth=shibb&url=http://search.ebscohost.com/login.aspx?direct=true&db=cat00846a&AN=ucfl.027732498&site=eds-live&scope=site

Horizon Research, Inc. (2017). National survey of science & mathematics education. Retrieved July 2, 2017, from http://horizon-research.com/NSSME/

Houseal, A., Gillis, V., Helmsing, M., & Hutchison, L. (2016). Disciplinary literacy through the lens of the Next Generation Science Standards. *Journal of Adolescent & Adult Literacy*, *59*(4), 377–384.

Huff, K., & Yager, R. E. (2016). The four strands of science learning and the Next Generation Science Standards. *Science Scope*, *40*(2), 10–13.

Hurd, P. D. (1958). Science literacy: Its meaning for American schools. *Educational Leadership*, *16*, 13–16,52.

Hurd, P. D. (1982). *An overview of science education in the United States and selected foreign countries.* (141 No. ED 227 076) (p. 128). Washington, DC: National Commission on Excellence in Education. Retrieved from https://eric.ed.gov/?id=ED227076

Hurd, P. D. (1986). Perspectives for the reform of science education. *The Phi Delta Kappan*, *67*(5), 353–358.

Hurd, P. D. (2000). *Transforming middle school science education*. New York, NY: Teachers College Press. Retrieved from https://login.ezproxy.net.ucf.edu/login?auth=shibb&url=http://search.ebscohost.com/login.aspx?direct=true&db=nlebk&AN=53645&site=eds-live&scope=site

Hurley, M. M. (2001). Reviewing integrated science and mathematics: The search for evidence and definitions from new perspectives. *School Science & Mathematics*, *101*(5), 259.

Huxley, T. H. (1881). *Science and culture*. England: Macmillan.

IBM Support. (2016). IBM SPSS origin and robustness of "equal variances not assumed" independent samples t-test [CT742]. Retrieved November 4, 2017, from http://www.ibm.com/support, //www.ibm.com/support/docview.wss?uid=swg21497421

International Association for the Evaluation of Educational Achievement. (2016). *Student achievement: TIMSS 2015 and TIMSS advanced 2015 international results*. Boston: Lynch School of Education, Boston College. Retrieved from http://timss2015.org/timss-2015/science/student-achievement/

International Association for the Evaluation of Educational Achievement. (2017). About TIMSS 2015. Retrieved September 9, 2017, from http://timss2015.org/timss-2015/about-timss-2015/

International Bureau of Education-UNESCO. (2017a). Discipline-based curriculum. *International Bureau of Education Glossary of Curriculum Terminology*. Geneva: UNESCO. Retrieved from http://www.ibe.unesco.org/en/glossary-curriculum-terminology/d/discipline-based-curriculum

International Bureau of Education-UNESCO. (2017b). Interdisciplinary approach. *International Bureau of Education Glossary of Curriculum Terminology*. Geneva: UNESCO. Retrieved from http://www.ibe.unesco.org/en/glossary-curriculum-terminology/d/discipline-based-curriculum

Jackson, J., & Ash, G. (2012). Science achievement for all: Improving science performance and closing achievement gaps. *Journal of Science Teacher Education*, *23*(7), 723–744. https://doi.org/10.1007/s10972-011-9238-z

Johnson, S. M., Kardos, S. M., Kauffman, D., Liu, E., & Donaldson, M. L. (2004). The support gap: New teachers' early experiences in high-income and low-income schools. *Education Policy Analysis Archives*, *12*(61), 1–7.

Jones, B. D. (2007). The unintended outcomes of high-stakes testing. *Journal of Applied School Psychology*, *23*(2), 65–86. https://doi.org/10.1300/J370v23n02_05

Jones, L. R., Wheeler, G., & Centurino, V. A. S. (2016). Chapter 2: TIMSS 2015 science framework. In M. Martin, I. V. S. Mullis, & M. Hooper (Eds.), *TIMSS 2015 Assessment Frameworks* (pp. 29–59). Chestnut Hill, MA: International Association for the Evaluation of Educational Achievement, TIMSS & PIRLS International Study Center, Lynch School of Education, Boston College, and International Association for the Evaluation of Education. Retrieved from http://timss.bc.edu/timss2015/downloads/T15_FW_Chap2.pdf

Kahana, M. J., Howard, M. W., & Polyn, S. M. (2008). Associative retrieval processes in episodic memory. *Psychology*, *Paper 3*.

Kahle, J. B., Meece, J., & Scantlebury, K. (2000). Urban African-American middle school science students: Does standards-based teaching make a difference? *Journal of Research in Science Teaching*, *37*(9), 1019–1041. https://doi.org/10.1002/1098-2736(200011)37:9<1019::AID-TEA9>3.0.CO;2-J

Kalyuga, S., Chandler, P., & Sweller, J. (1999). Managing split-attention and redundancy in multimedia instruction. *Applied Cognitive Psychology*, *13*, 351–371.

Kastberg, D., Chan, J. Y., & Murray, G. (2016). Performance of U.S. 15-year-old students in science, reading, and mathematics literacy in an international context: First look at PISA 2015. *National Center for Education Statistics*. Retrieved from https://eric.ed.gov/?id=ED570968

Kesidou, S., & Roseman, J. (2002). How well do middle school science programs measure up? Findings from Project 2061's curriculum review. *Journal of Research in Science Teaching*, *39*(6), 522–549.

Kieffer, M. J., Lesaux, N. K., Rivera, M., & Francis, D. J. (2009). Accommodations for English language learners taking large-scale assessments: A meta-analysis on effectiveness and validity. *Review of Educational Research*, *79*(3), 1168–1201. https://doi.org/10.3102/0034654309332490

Kim, J. S. (2007). Estimating item response theory models using Markov chain Monte Carlo methods. *Educational Measurement: Issues and Practices*, *38*(51).

Kohr, R. L., & Games, P. A. (1974). Robustness of the analysis of variance, the Welch procedure, and a box procedure to heterogeneous variances. *Journal of Experimental Education*, *43*(1), 61–69.

Koran, M. (2016). When it comes to school districts, size does matter. *Voice of San Diego*. Retrieved from http://www.voiceofsandiego.org/topics/education/when-it-comes-to-school-districts-size-does-matter/

Kozma, R. (2000). Reflections on the state of educational technology research and development. *Educational Technology Research and Development*, *48*(1), 5–15.

Krejcie, R. V., & Morgan, D. W. (1970). Determining sample size for research activities. *Education and Psychological Measurement*, *30*(3), 607–610.

Labaree, D. F. (2013). Let's measure what no one teaches: PISA, NCLB, and the shrinking aims of education. *Teachers College Record*, *116*(090303), 14.

Ladd, H. F. (2012). Education and poverty: Confronting the evidence. *Journal of Policy Analysis and Management*, *31*(2), 203–227. https://doi.org/10.1002/pam.21615

Lam, S., Wing-yi Cheng, R., & Choy, H. (2010). School support and teacher motivation to implement project-based learning. *Learning and Instruction*, *20*(6), 487–497. https://doi.org/10.1016/j.learninstruc.2009.07.003

Langdon, D., McKittrick, G., Beede, D., Khan, B., & Doms, M. (2011). *STEM: Good jobs now and for the future*. Washington, DC: US Department of Commerce.

Lara-Alecio, R., Tong, F., Irby, B. J., Guerrero, C., Huerta, M., & Fan, Y. (2012). The effect of an instructional intervention on middle school English learners' science and English reading achievement. *Journal of Research in Science Teaching*, *49*(8), 987–1011.

Lara-Cinisomo, S., Pebley, A. R., Valana, M. E., Maggio, E., Berends, M., & Lucas, S. R. (2004). *A matter of class: Educational achievement reflects family background more than ethnicity or immigration* (RAND Review). Santa Monica, CA: RAND Corporation. Retrieved from http://www.rand.org/publications/randreview/issues/fall2004/class.html

Lederman, N. G., & Niess, M. L. (1997). Integrated, interdisciplinary, or thematic instruction? Is this a question or is it questionable semantics? *School Science & Mathematics*, *97*(2), 57.

Lee, O. (2005). Science education with English language learners: Synthesis and research agenda. *Review of Educational Research*, *75*(4), 491–530.

Lee, O., & Buxton, C. A. (2013). Teacher professional development to improve science and literacy achievement of English language learners. *Theory into Practice*, *52*(2), 110–117.

Lepper, M. R., Greene, D., & Nisbett, R. E. (1973). Undermining children's intrinsic interest with extrinsic rewards: A test of the "overjustification hypothesis". *Journal of Personality and Social Psychology*, *28*(1), 129–137.

Lerner, L. S., Goddenough, U., Lynch, J., Schwartz, M., Schwartz, R., & Gross, P. R. (2012). *The state of state science standards*. Washington, DC: Thomas B. Fordham Institute. Retrieved from https://learningforward.org/docs/pdf/thestateofstatesciencestandards.pdf?sfvrsn=2

Leslie, L. L., McClure, G. T., & Oaxaca. (1998). Women and minorities in science and engineering: A life sequence analysis. *The Journal of Higher Education*, *69*(3), 239. https://doi.org/10.2307/2649188

Lippman, L., Burns, S., & McArthur, E. (1996). *Urban schools: The challenge of location and poverty* (No. NCES 96-184). National Center for Education Statistics. Retrieved from https://nces.ed.gov/pubs/web/96184ex.asp

Lovett, M. (2017). Learning principles-teaching excellence & educational innovation - Carnegie Mellon University. Retrieved December 17, 2017, from https://www.cmu.edu/teaching/principles/learning.html

Lunenburg, F. C., & Irby, B. J. (2008). *Writing a successful thesis or dissertation*. Thousand Oaks, CA: Corwin Press.

MacDonald, V.-M. (2004). The status of English language learners in Florida: Trends and prospects. Retrieved April 22, 2017, from http://epsl.asu.edu/epru/documents/EPSL-0401-113-EPRU.pdf

Maerten-Rivera, J., Ahn, S., Lanier, K., Diaz, J., & Lee, O. (2016). Effect of a multiyear intervention on science achievement of all students including English language learners. *Elementary School Journal*, *116*(4), 600–624.

Makashvili, M., & Slowinsky, E. (2009). *On the advantage of integrated science education in the middle school years*. Tbilisi, Georgia: National Curriculum & Assessment Centre, Ilia Chavchavadze University.

Marsh, H. W., Chanal, J. P., & Sarrazin, P. G. (2006). Self-belief does make a difference: A reciprocal effects model of the causal ordering of physical self-concept and gymnastics performance. *Journal of Sports Sciences*, *24*(1), 101–111. https://doi.org/10.1080/02640410500130920

Marsh, H. W., Trautwein, U., Ludtke, O., Koller, O., & Baumert, J. (2005). Academic self-concept, interest, grades, and standardized test scores: Reciprocal effects models of causal ordering. *Child Development*, *76*(2), 397.

Martin, M. O., Mullis, I. V. S., & Foy, P. (2016). Chapter 4: TIMSS 2015 assessment design. In M. O. Martin, I. V. S. Mullis, & M. Hooper (Eds.), *Methods and Procedures in TIMSS 2015* (pp. 85–99). Chestnut Hill, MA: International Association for the Evaluation of Educational Achievement, TIMSS & PIRLS International Study Center, Lynch School of Education, Boston College, and International Association for the Evaluation of Education. Retrieved from https://timssandpirls.bc.edu/publications/timss/2015-methods/chapter-1.html

Marx, R. W., & Harris, C. J. (2006). No Child Left Behind and science education: Opportunities, challenges, and risks. *The Elementary School Journal*, (5), 467. https://doi.org/10.1086/505441

Matkins, J., McDonnough, J., & Henschel, M. (2014). *Urban and suburban institutions: Preparing science teachers for culturally diverse classrooms*. Paper presented to National Association of Research in Science Teaching, Virginia Commonwealth University.

Mayer, R. E. (2008). Applying the science of learning: Evidence-based principles for the design of multimedia instruction. *The American Psychologist*, (8), 760.

Mayer, V. J. (Ed.). (2002). *Global science literacy*. Boston: Kluwer Academic.

McLaughlin, C. (2014). Urban science education: Examining current issues through a historical lens. *Cultural Studies of Science Education*, *9*(4), 885–923. https://doi.org/10.1007/s11422-014-9598-8

McMillen, B. J. (2004). School size, achievement, and achievement gaps. *Education Policy Analysis Archives*, *12*(58), 58.

Merritt, R. D. (2015). Integrated curriculum. *Research Starters: Education (Online Edition)*. Retrieved from https://login.ezproxy.net.ucf.edu/login?auth=shibb&url=http://search.ebscohost.com/login.aspx?direct=true&db=ers&AN=89164277&site=eds-live&scope=site

Metcalf, H. (2010). Stuck in the pipeline: A critical review of STEM workforce literature. *InterActions: UCLA Journal of Education and Information Studies*, *6*(2). Retrieved from http://escholarship.org/uc/item/6zf09176

Miller, P., Votruba-Drzal, E., & Setodji, C. M. (2013). Family income and early achievement across the urban-rural continuum. *Developmental Psychology*, *49*(8), 1452–1465.

Morgan, W. A. (1939). A test for the significance of the difference between the two variances in a sample from a normal bivariate population. *Biometrika*, *31*(1–2), 13–19.

Mullis, I. V. S. (2013). *TIMSS 2015 assessment frameworks*. Chestnut Hill, MA: International Association for the Evaluation of Educational Achievement, TIMSS & PIRLS International Study Center, Lynch School of Education, Boston College, and International Association for the Evaluation of Education.

Mullis, I. V. S., & Martin, M. O. (2013). *TIMSS 2015 item writing guidelines*. Chestnut Hill, MA: International Association for the Evaluation of Educational Achievement, TIMSS & PIRLS International Study Center, Lynch School of Education, Boston College, and International Association for the Evaluation of Education. Retrieved from http://timssandpirls.bc.edu/publications/timss/2015-methods/pdf/T15_item_writing_guidelines.pdf

National Academies of Science. (1996). *National science education standards*. Washington, DC: National Academies Press. https://doi.org/10.17226/4962

National Academy of Sciences. (2000). *Inquiry and the National Science Education Standards: A guide for teaching and learning*. Washington, DC: National Academies Press. Retrieved from https://login.ezproxy.net.ucf.edu/login?auth=shibb&url=http://search.ebscohost.com/login.aspx?direct=true&db=nlebk&AN=33528&site=eds-live&scope=site

National Academy of Sciences. (2007). *Rising above the gathering storm: Energizing and employing America for a brighter economic future*. Washington, DC: National Academies Press.

National Center for Education Statistics. (2006). Urban education in America: Definitions. Retrieved April 22, 2017, from https://nces.ed.gov/surveys/urbaned/definitions.asp

National Center for Education Statistics. (2007). *Comparing TIMSS with NAEP and PISA in Mathematics and Science*. National Center for Education Statistics, U.S. Department of Education.

National Center for Education Statistics. (2015a). NAEP - 2015 Science Assessment. Retrieved April 22, 2017, from https://nationsreportcard.gov/science_2015/#gaps?grade=8

National Center for Education Statistics. (2015b). NAEP Nations Report Card - Questionnaires for Students, Teachers, and Schools. National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education. Retrieved from https://nces.ed.gov/nationsreportcard/bgquest.aspx

National Center for Education Statistics. (2015c). *Program for International Student Assessment (PISA) - Overview*. Washington, DC: U.S. Department of Education, National Center for Education Statistics. Retrieved from https://nces.ed.gov/surveys/pisa/

National Center for Education Statistics. (2015d). Program for International Student Assessment (PISA) - Questionnaires. National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education. Retrieved from https://nces.ed.gov/surveys/pisa/questionnaire.asp

National Center for Education Statistics. (2015e). Trends in International Mathematics and Science Study (TIMSS) - Questionnaires. National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education. Retrieved from https://nces.ed.gov/timss/questionnaire.asp

National Center for Education Statistics. (2016). The condition of education 2016: English language learners in public schools. Retrieved April 22, 2017, from https://nces.ed.gov/fastfacts/display.asp?id=96

National Center for Education Statistics. (2017a). *Digest of education statistics, 2016*. Washington, DC: U.S. Department of Education. Retrieved from https://nces.ed.gov/programs/digest/d16/tables/dt16_214.10.asp?current=yes

National Center for Education Statistics. (2017b). *National assessment of educational progress: An overview of NAEP*. Washington, DC: National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education. Retrieved from https://nces.ed.gov/nationsreportcard/about/naephistory.aspx

National Center for Education Statistics. (2017c). *National assessment of educational progress: Science assessment*. Washington, DC: National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education. Retrieved from https://nationsreportcard.gov/science_2015/#scores?grade=8

National Center for Education Statistics. (2017d). *NCES common core of data: Locale codes*. Washington, DC: U.S. Department of Education, National Center for Education Statistics. Retrieved from https://nces.ed.gov/ccd/PDF/states/FL.pdf

National Center for Education Statistics. (2017e). *PISA data explorer*. Washington, DC: U.S. Department of Education, National Center for Education Statistics. Retrieved from https://nces.ed.gov/surveys/pisa/idepisa/report.aspx

National Center for Education Statistics. (2017f). *TIMSS data explorer*. Washington, DC: U.S. Department of Education, National Center for Education Statistics. Retrieved from https://nces.ed.gov/timss/timss2015/

National Council on Teacher Quality. (2017). *The all-purpose science teacher: An analysis of loopholes in state requirements for high school science teachers*. Washington, DC. Retrieved from http://www.nctq.org/dmsView/The_All_Purpose_Science_Teacher_NCTQ_Report

National Research Council. (1996). *National science education standards*. Washington, DC: National Academy Press. Retrieved from https://login.ezproxy.net.ucf.edu/login?auth=shibb&url=http://search.ebscohost.com/login.aspx?direct=true&db=cat00846a&AN=ucfl.021997724&site=eds-live&scope=site

National Research Council. (2012). *A framework for K-12 science education: Practices, crosscutting concepts, and core ideas*. Washington, DC: The National Academies Press. Retrieved from https://www.nap.edu/download/13165

National Science Board. (2007). *National action plan for addressing the critical needs of the U.S. science, technology, engineering, and mathematics education system*. Arlington, VA: National Science Foundation. Retrieved from https://www.nsf.gov/publications/pub_summ.jsp?ods_key=nsb07114

National Science Board. (2016). *Science and engineering indicators 2016* (No. NSB-2016-1). Arlington, VA: National Science Foundation. Retrieved from https://www.nsf.gov/statistics/2016/nsb20161/uploads/1/nsb20161.pdf

National Science Foundation. (2017). National Science Foundation: Funding. Retrieved July 2, 2017, from https://www.nsf.gov/funding/aboutfunding.jsp

National Science Foundation, National Center for Science and Engineering Statistics. (2016). *Science and engineering indicators 2016* (No. NSB-2016-1). Arlington, VA. Retrieved from https://www.nsf.gov/statistics/2016/nsb20161/#/

National Science Teachers Association. (2005). *Developing a world view for science education in North America and across the globe: Final report of the international task force*. Retrieved from http://www.nsta.org/pdfs/IntlTaskForceReport.pdf

National Science Teachers Association. (2016). NGSS hub. Retrieved July 1, 2017, from http://ngss.nsta.org/About.aspx

National Student Clearinghouse. (2015). *S&E degrees becoming more prevalent* (p. 43). Washington, DC: Lumina Foundation. Retrieved from https://eric.ed.gov/?id=ED558783

Neuenschwander, M. P., Vida, M., Garrett, J. L., & Eccles, J. S. (2007). Parents' expectations and students' achievement in two western nations. *International Journal of Behavioral Development*, *31*(6), 594–602. https://doi.org/10.1177/0165025407080589

NGSS Lead States. (2013a). Appendix D all standards, all students: Making the next generation science standards accessible to all students. In *Next Generation Science Standards: For states, by states* (Vol. 2, pp. 25–39). Washington, DC: National Academies Press. Retrieved from http://www.nextgenscience.org/sites/default/files/resource/files/Why%20K12%20Standards%20Matter.pdf

NGSS Lead States. (2013b). *Next generation science standards: For states, by states*. Washington, DC: The National Academies Press. Retrieved from http://nextgenscience.org/get-to-know

NGSS Lead States. (2013c). *Why K12 standards matter*. Washington, DC: National Research Council. Retrieved from http://www.nextgenscience.org/sites/default/files/resource/files/Why%20K12%20Standards%20Matter.pdf

NGSS Lead States. (2017). NGSS lead state partners. Retrieved October 7, 2017, from http://www.nextgenscience.org/lead-state-partners

Norton, A. O. (2001). Lectures on the Harvard Classics: Huxley on Science and Culture. In W. A. Neilson (Ed.), *The Harvard Classics, 1909-14* (Vol. XLI). New York, NY: P.F. Collier & Son. Retrieved from http://www.bartleby.com/60/185.html

Olson, S., & Riordan, D. G. (2012). *Report to the President, engage to excel: Producing one million additional college graduates with degrees in science, technology, engineering, and mathematics* (p. 103). Washington, DC: President's Council of Advisors on Science and Technology. Retrieved from https://eric.ed.gov/?id=ED541511

Onwuegbuzie, A. J. (2000). *Expanding the framework of internal and external validity in quantitative research*. Paper presented at the Annual Meeting of the Association for the Advancement of Educational Research, Ponte Vedra, FL. Retrieved from https://eric.ed.gov/?id=ED448205

Organization for Economic Cooperation and Development. (2016a). *Country note: Key findings from PISA 2015 for the United States* (p. 73). Paris: Organization for Economic Cooperation and Development. Retrieved from https://www.oecd.org/pisa/PISA-2015-United-States.pdf

Organization for Economic Cooperation and Development. (2016b). *PISA 2015 assessment and analytical framework: Science, reading, mathematic and financial literacy*. Paris: Organization for Economic Cooperation and Development. https://doi.org/10.1787/9789264255425-en

Organization for Economic Cooperation and Development. (2016c). *PISA 2015 results in focus* (p. 15). Paris: Organization for Economic Cooperation and Development. Retrieved from http://search.proquest.com/openview/e8c066902afa5f1a32b207d74370b87f/1?pq-origsite=gscholar&cbl=2026456

Organization for Economic Cooperation and Development. (2016d). *PISA in focus: How does PISA assess science literacy?* Paris: Organization for Economic Cooperation and Development. Retrieved from http://www.oecd-ilibrary.org/docserver/download/5jln4nfnqt7l-en.pdf?expires=1505250117&id=id&accname=guest&checksum=6656D065070A0AF0EC537568C76900AF

Organization for Economic Cooperation and Development. (2017). *About PISA*. Paris: Organization for Economic Cooperation and Development. Retrieved from http://www.oecd.org/pisa/aboutpisa/

Ornstein, A. C., & Levine, D. U. (2000). *Foundations of education*. Boston, MA: Houghton Mifflin.

Orr, C. (2008). *The ABCs of pattern scoring*. Tallahassee, FL: Florida Department of Education. Retrieved from fcat.fldoe.org/ppt/ABCsOfPatternScoring.ppt

Pajares, F. (1996). Self-efficacy beliefs in academic settings. *Review of Educational Research*, *66*(4), 543–578. https://doi.org/10.3102/00346543066004543

Pearson VUE. (2014). Test development and psychometric services. Retrieved August 13, 2017, from https://www.dhp.virginia.gov/dentistry/minutes/2017/ExamCte04282017Agenda.pdf

Pellegrino, J. W., Wilson, M. R., Koenig, J. A., & Beatty, A. S. (2014). *Developing assessments for the Next Generation Science Standards*. Washington, DC: National Academies Press.

Pitman, E. J. G. (1939). A note on normal correlation. *Biometrika*, *31*(1–2), 9–12.

Porter, A. C. (2002). Measuring the content of instruction: Uses in research and practice. *Educational Researcher*, *31*(7), 3–14.

Pringle, R. M., Mesa, J., & Hayes, L. (2017). Professional development for middle school science teachers: Does an educative curriculum make a difference? *Journal of Science Teacher Education*, *28*(1), 57.

Pursitasari, I. D., Nuryanti, S., & Rede, A. (2015). Promoting of thematic-based integrated science learning on the junior high school. *Journal of Education and Practice*, *6*(20), 97–101.

Ragel, R. (2015). Difference between discipline and subject. Retrieved July 1, 2017, from http://www.differencebetween.com/difference-between-discipline-and-vs-subject/

Reardon, S. F. (2013). The widening income achievement gap. *Educational Leadership*, *70*(8), 10–16.

Rich, J. T. (2017). Item response theory. Retrieved August 5, 2017, from http://www.psychologicaltesting.com/IRT.htm

Rigdon, E. E. (1996). CFI versus RMSEA: A comparison of two fit indexes for structural equation modeling. *Structural Equation Modeling*, *3*(4), 369–379.

Rivera Maulucci, M. S. (2010). Resisting the marginalization of science in an urban school: Coactivating social, cultural, material, and strategic resources. *Journal of Research in Science Teaching*, *47*(7), 840–860.

Rivera Maulucci, M. S., Brown, B. A., Grey, S. T., & Sullivan, S. (2014). Urban middle school students' reflections on authentic science inquiry. *Journal of Research in Science Teaching*, *51*(9), 1119–1149. https://doi.org/10.1002/tea.21167

Roediger, H. L. I. (2006). The power of testing memory: Basic research and implications for educational practice. *Psychological Science*, *1*(3), 181–210.

Roehrig, G., Kruse, R., & Kern, A. (2007). Teacher and school characteristics and their influence on curriculum implementation. *Journal of Research in Science Teaching*, *44*(7), 883–907.

Roth, W.-M., & Barton, A. C. (2004). *Rethinking scientific literacy*. New York, NY: Routledge. Retrieved from https://login.ezproxy.net.ucf.edu/login?auth=shibb&url=http://search.ebscohost.com/login.aspx?direct=true&db=nlebk&AN=102980&site=eds-live&scope=site

Ruby, A. (2006). Improving science achievement at high-poverty urban middle schools. *Science Education*, *90*(6), 1005–1027.

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

Sahlberg, P., & Hargreaves, A. (2011). *Finnish lessons: What can the world learn from educational change in Finland?* New York, NY: Teachers College Press.

Salzman, H. (2013). What shortages? The real evidence about the STEM workforce. *Issues in Science and Technology*, *29*(4), 58–67.

Santau, A. O., Secada, W., Maerten-Rivera, J., Cone, N., & Lee, O. (2010). U.S. urban elementary teachers' knowledge and practices in teaching science to English language learners: Results from the first year of a professional development intervention. *International Journal of Science Education*, *32*(15), 2007.

Sass, T. R. (2013). *The market for new science teachers in Florida: A report to the National Research Council Committee on strengthening science education through a teacher learning continuum*. Washington, DC: National Research Council.

Sawilowsky, S. S. (2003). A different future for social and behavioral science research. *Journal of Modern Applied Statistical Methods*, *2*(1), 128–132. https://doi.org/10.22237/jmasm/1051747860

Schaffer, C. L., White, M., & Brown, C. M. (2016). *Questioning assumptions and challenging perceptions: Becoming an effective teacher in urban environments*. Lanham, MD: Rowman & Littlefield.

Schindel Dimick, A. (2016). Exploring the potential and complexity of a critical pedagogy of place in urban science education. *Science Education*, *100*(5), 814–836.

Schmider, E., Ziegler, M., Danay, E., Beyer, L., & Bühner, M. (2010). Is it really robust? *Methodology*, *6*(4), 147–151. https://doi.org/10.1027/1614-2241/a000016

Scott, E. (2016). *Comparing NAEP, TIMSS, and PISA in mathematics and science*. Washington, DC: National Center for Education Statistics, Institute of Education Sciences, U.S. Dept. of Education. Retrieved from https://nces.ed.gov/timss/pdf/naep_timss_pisa_comp.pdf

Scott, E. C., & Branch, G. (2008). Anti-evolution legislation in the Bayou State. *Reports of the National Center for Science Education*, *28*(2), 8–11.

Shamos, M. H. (1995). *The myth of scientific literacy*. New Brunswick, N.J.: Rutgers University Press. Retrieved from https://login.ezproxy.net.ucf.edu/login?auth=shibb&url=http://search.ebscohost.com/login.aspx?direct=true&db=cat00846a&AN=ucfl.020290195&site=eds-live&scope=site

Sherriff, R. (2014). Middle school madness: The integrated or discipline specific choice. Retrieved May 14, 2017, from http://www.classroomscience.org/middle-school-madness

Sherriff, R. (2015). Middle school madness part 2: Integrated science versus coordinated science. Retrieved June 7, 2017, from http://www.classroomscience.org/middle-school-madness-part-2-integrated-science-versus-coordinated-science

Silver, H. F., Strong, R. W., & Perini, M. J. (2000). *So each may learn: Integrating learning styles and multiple intelligences*. Alexandria, VA: Association for Supervision and Curriculum Development.

Slovacek, S., Whittinghill, J., Flenoury, L., & Wiseman, D. (2012). Promoting minority success in the sciences: The minority opportunities in research programs at CSULA. *Journal of Research in Science Teaching*, *49*(2), 199.

Steinberg, W. J. (2011). *Statistics alive* (2nd ed.). Los Angeles: SAGE Publications, Inc.

Stengel, B. S. (1997). "Academic discipline" and "school subject": Contestable curricular concepts. *Journal of Curriculum Studies*, *29*(5), 585–602. https://doi.org/10.1080/002202797183928

Stoddart, T., Pinal, A., Latzke, M., & Canaday, D. (2002). Integrating inquiry science and language development for English language learners. *Journal of Research in Science Teaching*, *39*(8), 664–687. https://doi.org/10.1002/tea.10040

Stoica, I. (2015). Curriculum theory. *Research Starters: Education (Online Edition)*. Retrieved from https://login.ezproxy.net.ucf.edu/login?auth=shibb&url=http://search.ebscohost.com/login.aspx?direct=true&db=ers&AN=89164157&site=eds-live&scope=site

Strauss, V. (2013). A debate: What do international test scores tell us? *Washington Post*. Retrieved from https://www.washingtonpost.com/news/answer-sheet/wp/2013/11/20/a-debate-what-do-international-test-scores-tell-us/

Student Assessment Division. (2008). *A study of the correlation between grade 10 math performance and course performance* (p. 6). Austin, TX: Texas Education Agency.

Sturges, A. W. (1976). Forces influencing the curriculum. *Educational Leadership*, *34*(1), 40–43.

Tal, T., Krajcik, J. S., & Blumenfeld, P. C. (2006). Urban schools' teachers enacting project-based science. *Journal of Research in Science Teaching*, *43*(7), 722–745. https://doi.org/10.1002/tea.20102

Tamassia, L., & Frans, R. (2014). Does integrated science education improve scientific literacy? *Journal of the European Teacher Education Network*, *9*, 131–141.

Tan, E., & Barton, A. C. (2010). Transforming science learning and student participation in sixth grade science: A case study of a low-income, urban, racial minority classroom. *Equity & Excellence in Education*, *43*(1), 38–55. https://doi.org/10.1080/10665680903472367

Tanner, D., & Tanner, L. N. (1990). *History of the school curriculum*. New York, NY: Macmillan.

Taylor, G. S., & Hord, C. (2016). An exploratory analysis of a middle school science curriculum: Implications for students with learning disabilities. *Learning Disabilities: A Multidisciplinary Journal*, *21*(2), 1–13. https://doi.org/10.18666/LDMJ-2016-V21-I2-7582

Thadani, V., Cook, M. S., Griffis, K., Wise, J. A., & Blakey, A. (2010). The possibilities and limitations of curriculum-based science inquiry interventions for challenging the "pedagogy of poverty." *Equity & Excellence in Education*, *43*(1), 21–37. https://doi.org/10.1080/10665680903408908

Tobin, K. (1986). Validating teacher performance measures against student engagement and achievement in middle school science. *Science Education*, *70*(5), 539.

Trevino Jr, D., Braley, R. T., Brown, M. S., & Slate, J. R. (2008). Challenges of the public school superintendency: Differences by tenure and district location. *Florida Journal of Educational Administration & Policy*, *1*(2), 98–109.

Tucker, L. R., & Lewis, C. (1973). A reliability coefficient for maximum likelihood factor analysis. *Psychometrika*, *38*, 1–10.

Tucker, M. (Ed.). (2011). *Surpassing Shanghai: An agenda for American education built on the world's leading systems*. Cambridge, MA: Harvard Education Press. Retrieved from http://hepg.org/hep-home/books/surpassing-shanghai

Tyler, R. W. (1950). *Basic principles of curriculum and instruction*. Chicago, IL: Syllabus Division, University of Chicago Press.

U.S. Census Bureau. (2014). *Majority of STEM college graduates do not work in STEM occupations* (News Release No. CB14-130). Washington, DC: U.S. Department of Commerce. Retrieved from https://www.census.gov/newsroom/press-releases/2014/cb14-130.html

U.S. Department of Education. (2015). *NAEP - 2015 Science Assessment*. Washington, DC: Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP). Retrieved from https://www.nationsreportcard.gov/science_2015/#scores?grade=8

U.S. Department of Education. (2016a). Progress in our schools. Retrieved March 16, 2017, from https://www.ed.gov/k-12reforms?src=rn

U.S. Department of Education. (2016b). Science, technology, engineering and math: Education for global leadership. Retrieved March 15, 2017, from https://www.ed.gov/Stem

Venville, G. J., Wallace, J., Renne, L. J., & Malone, J. A. (2002). Curriculum integration: Eroding the high ground of science as a school subject? *Studies in Science Education*, *37*, 43–83.

Visone, J. D. (2009). The validity of standardized testing in science. *American Secondary Education*, *38*(1), 46–61.

Visone, J. D. (2010). Science or reading: What is being measured by standardized tests? *American Secondary Education*, *39*(1), 95–112.

Vockley, M., & Lang, V. (2009). *Alignment and the states: Three approaches to aligning the national assessment of educational progress with state assessments, other assessments, and standards*. Council of Chief State School Officers. Retrieved from https://eric.ed.gov/?id=ED528641#?

Vogler, K. E., & Virtue, D. (2007). “Just the facts, ma’am”: Teaching social studies in the era of standards and high-stakes testing. *The Social Studies*, *98*(2), 54.

Waters, T., Marzano, R. J., & McNulty, B. (2003). *Balanced leadership: What 30 years of research tells us about the effect of leadership on student achievement*. Aurora, CO: Mid-Continent Research for Education and Learning. Retrieved from https://eric.ed.gov/?id=ED481972

Webb, N. L. (1997). *Criteria for alignment of expectations and assessments in mathematics and science education.* (Council of Chief State School Officers and National Institute for Science Education Research Monograph No. 6). Washington, DC.

Wenning, C. J. (2007). Scientific literacy: The main goal of science education. In *Using a modified Delphi Technique to validate a physics and physical science teacher education textbook* (pp. 24–37). Chicago, IL: Illinois State University. Retrieved from http://search.proquest.com/openview/b2aedb23c9e620f5e285365981f009b2/1?pq-origsite=gscholar&cbl=18750&diss=y

WestEd, & Council of Chief State School Officers. (2014). *Science Framework for the 2015 National Assessment of Educational Progress*. Washington, DC: National Assessment Governing Board, U.S. Department of Education. Retrieved from https://www.nagb.gov/naep-frameworks/science/2015-science-framework.html

White, M., Brown, C. M., Viator, M. G., Byrne, L. L., & Ricchezza, L. C. (2017). Transforming perceptions of urban education: Lessons from Rowan University's urban teacher academy. *Educational Forum*, *81*(1), 18–34. https://doi.org/10.1080/00131725.2016.1243181

Wilson, D. M., Bates, R., Scott, E. P., Painter, S. M., & Shaffer, J. (2015). Differences in self-efficacy among women and minorities in STEM. *Journal of Women and Minorities in Science and Engineering*, *21*(1). https://doi.org/10.1615/JWomenMinorScienEng.2014005111

Wilson, M. R., & Bertenthal, M. W. (2006). *Systems for state science assessment*. Washington, DC: National Academies Press. Retrieved from https://login.ezproxy.net.ucf.edu/login?auth=shibb&url=http://search.ebscohost.com/login.aspx?direct=true&db=nlebk&AN=148936&site=eds-live&scope=site

Wiseman, A. W. (2012). The impact of student poverty on science teaching and learning: A cross-national comparison of the South African case. *American Behavioral Scientist*, *56*(7), 941–960. https://doi.org/10.1177/0002764211408861

Wong, D., & Pugh, K. (2001). Learning science: A Deweyan perspective. *Journal of Research in Science Teaching*, *38*(3), 317–336.

Woods, J. R. (2016). *Mitigating teacher shortages: Alternative certification* (Teacher shortages: What we know). Washington, DC: Education Commission of the States. Retrieved from https://www.ecs.org/wp-content/uploads/Mitigating-Teacher-Shortages-Alternative-Certification.pdf

Wright, E., & Sunal, D. W. (2006). *The impact of state and national standards on K-12 science teaching*. Greenwich, CT: Information Age Publishing.

Xue, Y., & Larson, R. (2015). STEM crisis or STEM surplus? Yes and yes. *Monthly Labor Review*. Retrieved from https://doi.org/10.21916/mlr.2015.14.

Yetter, I. H., Livengood, K. K., & Smith, W. S. (2017). State science standards and students' knowledge of what states value: Lunar phases. *Electronic Journal of Science Education*, *21*(1), 36–55.

Youmans, E. L. (1867). *Modern culture; its true aims and requirements; A series of addresses and arguments on the claims of scientific education*. London: Macmillan and Co.

Zhbanova, K. S., Rule, A. C., Montgomery, S. E., & Nielsen, L. E. (2010). Defining the difference: Comparing integrated and traditional single-subject lessons. *Early Childhood Education Journal*, *38*(4), 251–258. https://doi.org/10.1007/s10643-010-0405-1