SHIFTING THE FOCUS TO SCIENCE IN THE EARLY ELEMENTARY YEARS: AN EXAMINATION OF SCIENCE ACHIEVEMENT GROWTH IN GRADES K-2 USING A NATIONALLY REPRESENTATIVE DATASET

by

TRACY E. BOUSSELOT

A DISSERTATION

Presented to the Department of Educational Methodology, Policy, and Leadership
and the Graduate School of the University of Oregon
in partial fulfillment of the requirements
for the degree of
Doctor of Philosophy

June 2018

DISSERTATION APPROVAL PAGE

Student: Tracy E. Bousselot

Title: Shifting the Focus to Science in the Early Elementary Years: An Examination of Science Achievement Growth in Grades K – 2 Using a Nationally Representative Dataset

This dissertation has been accepted and approved in partial fulfillment of the requirements for the Doctor of Philosophy degree in the Department of Educational Methodology, Policy, and Leadership by:

Gina Biancarosa	Chairperson
Joseph Stevens	Core Member
Ilana Umansky	Core Member
Joanna Goode	Institutional Representative

and

Sara D. Hodges	Interim Vice Provost and Dean of the Graduate School

Original approval signatures are on file with the University of Oregon Graduate School.

Degree awarded June 2018

DISSERTATION ABSTRACT

Tracy E. Bousselot

Doctor of Philosophy

Department of Educational Methodology, Policy, and Leadership

June 2018

Title: Shifting the Focus to Science in the Early Elementary Years: An Examination of Science Achievement Growth in Grades K – 2 Using a Nationally Representative Dataset

Efforts to understand growth and disparities in science achievement have mainly been focused on the middle and high school grades in studies of K – 12 science education, leaving a gap in the research about the early elementary years. This study used a nationally-representative sample of students in Grades K – 2 to examine science achievement and growth trajectories of students by gender and race/ethnicity. Using multilevel growth modeling, differences in science achievement at Grade 2 and in rate of growth were detected for several student groups. Socioeconomic status, prior reading and math achievement, and student home language status were also significant predictors of science achievement. Growth effect size estimates were calculated by student group and showed substantial year-to-year growth in science achievement in the early elementary grades, with a slight decrease in effect size across years. In order to strengthen current efforts to increase student engagement and participation in science and STEM-related career and college pathways, especially for historically underrepresented groups, policymakers should shift focus to better understand promising practices that best support all students in science from the onset of their K – 12 educational experience.

CURRICULUM VITAE

NAME OF AUTHOR: Tracy E. Bousselot

GRADUATE AND UNDERGRADUATE SCHOOLS ATTENDED:

University of Oregon, Eugene, OR
Western Illinois University, Macomb, IL

DEGREES AWARDED:

Doctor of Philosophy, Educational Leadership, 2018, University of Oregon
Master of Education in Educational Leadership, 2002, University of Oregon
Bachelor of Science, Biology, 1994, Western Illinois University

AREAS OF SPECIAL INTEREST:

K-12 Science Education
Science Literacy
STEM Professional Development
Program Evaluation

PROFESSIONAL EXPERIENCE:

Research Associate, Inflexion, 2013-Present
Research Assistant, Center on Teaching and Learning, 2009-2013
High School Teacher, Springfield Public Schools, 2006-2009
Middle School Teacher, South Lane School District, 2003-2005

PUBLICATIONS:

Breslow, J., Bousselot, T., & Chadwick, K. (2016). *Oregon Quality Education Commission: College Readiness Case Study Project*. Eugene, OR: Educational Policy Improvement Center.

Baker, D.L., Biancarosa, G., Park, B., Bousselot, T., Smith, J.L., Baker, S.K., Kame'enui, E.J., Alonzo, J., & Tindal, G. (2014). Validity of CBM measures of oral reading fluency and reading comprehension on high-stakes reading assessments in grades 7 and 8. *Reading and Writing: An Interdisciplinary Journal*, *27*, 1 – 48.

Cummings, K.D., Bousselot, T., Smith, J.M., Brown, M., Beemer, N., Kennedy, P.C., Munir-McHill, S., Baker, S.K., & Kame'enui, E.J. (2011). *Sentinel Schools in the DIBELS Data System (2010-2011): A Description of Methods* (Technical Report 1103). Eugene, OR: University of Oregon.

ACKNOWLEDGMENTS

I wish to thank all of my committee members, including Dr. Joanna Goode and Dr. Ilana Umansky, for agreeing to serve on my committee and for their support and feedback throughout the process. Additionally, I would be remiss for not expressing my sincere gratitude to Dr. Joe Stevens for telling me, at one point, to “dig in and hang on” through the rough parts, and to Dr. Gina Biancarosa, whose steadfast encouragement, guidance, and good cheer helped to keep me moving forward. I feel immensely privileged to have worked with such a talented and dedicated group of educators who challenged me to do my best with a topic about which I am most passionate.

My work is dedicated to my greatest influences – Fea Lucia, Ana Maria, and Ina Naia – the stars that keep me on course. Thank you to Lisa, Julia, Stacy, Alena, Erin, Holle, Christine, and Lori for inspiration, affirmation, and for constantly reminding me what truth, strength, integrity, and family look like. And to Keith, thank you for always bringing the chocolate.

TABLE OF CONTENTS

Chapter	Page
I. INTRODUCTION	1
II. LITERATURE REVIEW	7
Building a STEM Workforce	7
Science Literacy	9
Group Differences	11
Gender	11
Race/Ethnicity	13
Socioeconomic Status	14
Reading and Math Achievement	15
Growth Trajectories	16
School Level Factors	17
Previous Research on Science Achievement Using ECLS Data	21
Purpose of Study	22
Research Questions	23
III. METHODS	24
Study Data	24
Participants and Procedures	25
Assessment Administration	26
ECLS-K Measures	28
IRT-based Scores……	28
Reading Achievement Measure	29

Chapter	Page

Math Achievement Measure 29

Science Achievement Measure 30

Assessment Psychometric Data 31

General Classroom Teacher Questionnaires 31

Study Variables 31

Dependent Variable…… 31

Demographic Predictor Variables 31

Initial Reading and Math Scores 33

Science Instructional Time 33

Interaction Effects 34

Analytic Methods 34

Effect Size 37

Sample Weights 38

Missing Data 39

IV. RESULTS 40

Testing Assumptions 40

Missingness Analysis 42

Planned Missingness 45

Multiple Imputation 47

Design Effects 48

Analytic Sample 48

Descriptive Statistics 49

Chapter Page

Model Estimation 53

Unconditional Models 54

Conditional Models 57

Effect Size 65

V. DISCUSSION 67

Growth Trajectories 67

Demographic Predictor Variables 69

Prior Reading and Math Achievement 74

Science Instructional Time 76

Science Achievement Growth Effect Size 77

Limitations 79

Conclusion 80

APPENDIX: HISTOGRAMS AND SCATTERPLOTS OF SCIENCE ACHIEVEMENT SCORES ACROSS ASSESSMENT TIME POINTS 82

REFERENCES CITED 83

LIST OF FIGURES

Figure Page

1. Random sample of 25 science achievement growth trajectories, Grades K – 2 42

2. Estimated science achievement growth trajectories by gender................................ 60

3. Estimated science achievement growth trajectories by race/ethnicity..................... 61

LIST OF TABLES

Table Page

1. Approximate sample sizes (Rounded to the nearest 10) of student participants in the direct child assessments in the ECLS-K: 2011, K – 2 26

2. Reported reliabilities of IRT theta scores for reading, math, and science achievement assessments used in the ECLS-K: 2011, by data collection time point 30

3. Pairwise correlations among study variables 41

4. Number and percent of missing data by variable of interest 45

5. Weighted descriptive statistics for key predictor variables 49

6. Weighted science theta score means and standard deviations by student groups across grades 51

7. Random effects for unconditional longitudinal HLM regression models for alternate time coding schemes, grades K – 2 56

8. Fixed effects for longitudinal HLM regression models, grades K – 2 62

9. Random effects for longitudinal HLM regression models, grades K – 2 63

10. Fixed and random effects for longitudinal HLM regression model including instructional time and interaction terms, grades K – 2 64

11. Science achievement growth rate at each assessment time point by student group and annual growth rate effect sizes.. 66

CHAPTER I

INTRODUCTION

The popular anecdote often cited at the outset of calls to reform science education in the United States revolves around the launch of a shiny sphere of metal 58 cm in diameter in October of 1957. For three months, the sphere orbited the Earth before burning up on re-entry into the atmosphere in January 1958. During the time that the satellite, named the Sputnik 1, was in orbit, providing scientists with data on the density of the upper atmosphere, it could be seen from Earth as it circled overhead every 96 minutes. The effects of the visible and short-lived journey across the sky of this simple but technologically and historically significant device were to directly ignite the "Space Race" and, more peripherally, to open the door to a new era of science education reform in the United States (Clowse, 1981). The specific urgency that resulted from the successful travels of Sputnik centered on the fact that the satellite was designed and launched by the Soviet Union, which both scared and shocked citizens and policy makers in the United States. Fearing that the U.S. would lose its competitive edge and be outpaced by the Soviet Union in the areas of science, technology, engineering, and mathematics, the National Defense Education Act (NDEA) was passed by Congress in 1958. The NDEA was charged with providing funding designed to mobilize awareness of and stimulate reform in science education. At the same time, funding was increased for the National Science Foundation, which was already in existence. In addition, the National Aeronautics and Space Association (NASA) was established a year later.

The Space Race ended after the lunar landing of the Apollo 11 by U.S. astronauts and the decreasing interest in space exploration of the American public in the early

1970s; the emphasis on science education lost momentum in the ensuing years until the release of the controversial report *A Nation at Risk* in 1983. While the report has been criticized for, among other things, the incorrect presentation of contradictory data that was used to characterize U.S. schools as failing, one outcome of the report was a renewed interest in the sciences as attention was shifted to the increasing role of science and technology in a rapidly changing global economy. During the middle 80s to middle 90s, the acronym STEM was coined as a way to link the disciplines of science, technology, engineering, and mathematics, which, as the world headed toward the turn of the century, became seen as critical components to driving innovation and critical thinking in the next generation of U.S. students (Breiner, Harkness, Johnson, & Koehler, 2012).

However, in 2001, the advent of the No Child Left Behind Act (NCLB) resulted in educational efforts that laid out specific requirements for schools to meet adequate yearly progress under accountability rules that emphasized the importance of increasing scores on standardized tests of reading and mathematics. Science curriculum was reduced during the first six years of the enactment of NCLB to focus on reading and math (Cavanaugh, 2007). As the pendulum swung back toward a de-emphasis of science education once again, reports like *Before It's Too Late* (National Academy of Sciences, National Academy of Engineering, and Institute of Medicine, 2000) and *Rising Above the Gathering Storm* (Augustine, 2005) were released. The report authors were attempting to call attention to the fact that the nation's schools were still lagging in attempts to boost science education reform efforts and engage and retain student interest in careers in the STEM fields. Indeed, the original members of the committee that produced the original 2005 report of *Rising Above the Gathering Storm* were asked to reconvene and reexamine their original work by members of the government, resulting in the darkly humorous titled *Rising Above the Gathering Storm, Revisited: Rapidly Approaching Category 5* (Augustine et al., 2010). While the authors of the new report noted that there were immediate efforts in the U.S. to address the recommendations outlined in the original report, their summative assessment of conditions was that "…overall the United States long-term competitiveness outlook (read jobs) has further deteriorated since the publication of the *Gathering Storm* report five years ago" (p. 65). Augustine and colleagues (2010) pointed out that the need to invigorate science and math education in public schools was not just to increase the pool of candidates for the STEM workforce, but also because of the expansion of a technology-based workplace where basic skills in science and math are required regardless of job entry level.

The replacement of NCLB with the new accountability program known as the *Every Student Succeeds Act* (ESSA) went into effect at the beginning of the 2017-18 school year. Under this act, statewide science testing continues to occur once in each of the following grade spans: 3-5, 6-9 and 10-12 (ESSA, 2015). Advocates of the program point to ESSA as a step in the right direction away from the prescriptive provisions of NCLB that will allow for more flexibility in making curricular and testing decisions at the state level, but it is too soon to tell how ESSA may directly affect science education reform efforts.

Besides a plethora of reports and briefs signaling concern about science education in the U.S., a reinvigorated national emphasis on education in the STEM fields has been led by and brought to public and policy attention with the creation and adoption of the Next Generation Science Standards (NGSS; NGSS Lead States, 2013) by 19 states and the District of Columbia (as of November 2017). The NGSS serve as an update and replacement of the *National Science Education Standards* (NRC, 1996) and *Benchmarks for Science* (AAAS, 1994) that were used in creating state level science standards and have been heralded as a further call from multi-sector stakeholders to take seriously the need to increase student interest and achievement in the STEM disciplines. The intent of the NGSS is to refocus K–12 science education using multidisciplinary connections, an emphasis on hands-on and applied learning, and to highlight the importance of engineering and technology by integrating them more thoroughly into the science standards. The goal of implementation of the NGSS is to improve outcomes for students around college preparation, STEM career readiness, and the ability of all members of society to make informed decisions related to science literacy.

The crux of the urgency around the importance of finding ways for students to engage deeply with STEM fields and opportunities is this: In order for the U.S. to remain competitive and innovative in the fields of science, engineering, and technology, a new generation of students will need to develop the requisite science knowledge and literacy skills, including facility with reading, writing, and speaking the language of science (National Academy of Sciences, National Academy of Engineering, and Institute of Medicine, 2005). Beyond just competing in the current global economy, future generations of U.S. students will need to be able to contribute to future discoveries and challenges that humanity will face. Unfortunately, persistent differences in science achievement, attitude, and career choices, described in part by complex relationships between gender, race/ethnicity, and socioeconomic status, present challenges to fulfilling these goals.

Most of the existing literature on science achievement has been focused on older student populations, presumably because high school is where the specialization for a future in STEM majors/careers begins. Research has shown that high school students' achievement in STEM is predictive of whether students choose and persist in a STEM field in college (Berryman, 1983; Conley, 2007; Wang, 2013). High school STEM achievement has also been shown to explain both gender and racial gaps in the selection of a college or career path in the STEM disciplines (Miller & Kimmel, 2012; Marsh, Trautwein, Lüdtke, Köller, & Baumert, 2005). However, less is known about the effects of students' early experiences in the sciences, especially as to how growth and trajectories in science achievement may emerge and develop as students progress through elementary and middle school and how student level factors may affect science achievement gaps, as well as other school level contextual factors. A review of the literature by Byrnes and Miller in 2006 found only 12 longitudinal studies of science achievement in the 15 years prior to their work (Byrnes & Miller, 2006). The present study contributes to a lacuna in the research on science achievement, particularly in the area of longitudinal studies. The study explores the science achievement growth over time of a cohort of elementary school students from kindergarten through Grade 2 who participated in the Early Childhood Longitudinal Study (ECLS-K) from 2010 – 2015. The ECLS-K: 2011 study is unique in that it included a measure of science achievement starting in kindergarten, where most large scale longitudinal studies have focused on reading and math achievement and growth, affording a better understanding of where and how to focus attention when looking to improve opportunities for all students in science. Beyond just examining science achievement growth, data collected as part of the ECLS-

K: 2011 allow for the inclusion of factors that may underlie patterns of growth and corresponding achievement gaps, giving rise to potential guidance around promising practices for science instruction in the early grades.

CHAPTER II

LITERATURE REVIEW

Building a STEM Workforce

The effects of globalization and the demands of the 21st century workplace dictate a more sophisticated set of skills necessary for future workers as we enter an era with an increasingly competitive international work force (NRC, 2011). In a rapidly changing, technologically rich, global society, literacy in the fields of Science, Technology, Engineering, and Mathematics (STEM) is required to participate in, and drive, an innovation-based economy. Jobs in the 21st century require people with the knowledge, skills, and mindsets that will enable them to adapt to flexible workforce needs and to compete for high-wage, high-demand careers. It is widely recognized that high quality, cross-disciplinary STEM education encourages skills such as critical thinking, problem-solving, collaboration, and creativity (Rothwell, 2013). Employment projections by the U.S. Department of Labor show that 15 of the 20 fastest growing occupations projected for 2017 require significant mathematics or science preparation.

STEM fields are considered vital to the health of the U.S. economy (Atkinson & Mayo, 2010). Jobs in the STEM fields can potentially bring high economic rewards at both the individual and societal labor market levels (Rothwell, 2013). Amidst claims that the U.S. is losing its competitive edge in the global economy (National Academy of Science et al., 2007), organizations such as the National Research Council and a coalition of the National Academy of Science, National Academy of Engineering, and Institute for Medicine have released reports based on current trends and economic forecasts calling for significant reforms to K-12 science education in the U.S. to help address this growing need (NRC, 2011; National Academy of Sciences, National Academy of Engineering, and Institute of Medicine, 2005). The President's Council of Advisors on Science and Technology (2012) specifically laid out a goal to increase the number of STEM bachelor's degrees by 33% annually. Reports such as these are often used to provide support for STEM education initiatives, citing the need to build STEM interest and engagement at the local level in order to develop a home grown highly skilled work force ready to fill regional positions in the STEM fields. Not only is it important to maintain job growth in the STEM fields that is currently exceeding that of the non-STEM fields, workers in STEM occupations on average have higher earnings and experience lower unemployment rates over time than workers in other fields (Langdon, McKittrick, Beede, Khan, & Doms, 2011).

However, persistent achievement gaps in science education and degree attainment likely continue to contribute to disparities in the STEM workforce. While women make up half of all college-educated workers in the United States, they represented only 28% of science and engineering workers in 2010. Similarly, historically underrepresented racial and ethnic groups (i.e., Hispanic, Black, and American Indian or Alaska Natives) accounted for only 10% of the country's workers in science and engineering in 2010, quite low compared to their representation in the population of the U.S. (NSB, 2016). Despite efforts by stakeholders to increase student engagement with STEM-related pathways over the last several years, a recent report by the ACT signals that interest in pursuing a STEM major or occupation has remained virtually unchanged (ACT, 2018). Since 2012, less than half of the students who have taken the ACT college entrance test have indicated an intention to either follow a STEM-related major or occupation or indicated participation in activities that are aligned to STEM interests. Additionally, the report makes clear that differences in interest in and readiness for STEM pathways after high school continue by gender and underserved student populations as recently as the graduating class of 2017.

Science Literacy

Having an understanding of the basic tenets behind science and how it works should not be limited to those who enter STEM or STEM-related careers. A recent report by the National Academies of Sciences, Engineering, and Medicine (2016; NASEM) emphasizes the importance of and distinction between science literacy at the individual, community, and societal levels. The authors of the report define science literacy as a "…level of familiarity with the enterprise and practice of science" (p. 1). While content specific knowledge is one facet of science literacy, the broader picture involves a fundamental understanding of the process of scientific practice and decision-making. Increasingly, people are required to sort through large amounts of information and make decisions using a mix of STEM-related knowledge and skills. These decisions may be related to such topics as medical care or understanding the effects of climate change on the economy. Without the ability to locate and evaluate accurate and reliable information, today's students may have difficulty understanding current issues in science (Julien & Barker, 2009), among others. The arguments for the importance of increasing the general scientific and technological literacy of all citizens have been well articulated by multiple national organizations and advancement groups for decades (e.g., AAAS, 1991; NAE and NRC 2002; NRC 1989). With the continued advances in science and technology, it becomes more important to establish a nation of scientifically literate citizens capable of making informed decisions. However, research has shown a connection between the opportunity to access high-quality science education and the development of science literacy (NASEM, 2016), which may tend to further exacerbate differences based on social inequity.

The results of international comparisons and national achievement assessments present an unfavorable picture of the current state of science achievement in the U.S. The most recent report released by the Program for International Student Assessment shows U.S. science students well behind their global counterparts in science literacy, ranked at 19th out of 65 participating countries (PISA, 2012). The Trends in International Mathematics and Science Study (TIMSS; Kastberg, Roey, Ferraro, Lemanski, & Erberber, 2013), which every four years reports on the state of mathematics and science achievement for students from around the world, released science assessment results in 2011 that ranked U.S. fourth graders and eighth graders at 7th and 13th, respectively, out of the 63 countries and 14 benchmarking entities that participated in the study (TIMSS, 2011). It should be noted that not all experts agree on the usefulness of international comparisons to make judgments about the quality of science education in the U.S., and that quite often media reports citing the results of these assessments may paint a direr picture than warranted (e,g,. Bracey, 2000; Koretz, 2008). Additionally, the exclusion of students with disabilities from the large-scale international assessments like the TIMSS and the PISA give rise to questions about the overall generalizability and comparability of results as well as policy making decisions that might arise due to such comparisons (Schuelka, 2013).

International comparisons notwithstanding, most of our understanding of student achievement trends in K – 12 science comes from the National Assessment for Educational Progress (NAEP). The most recent results of the NAEP science assessment in 2015 showed evidence of improvement in science achievement for some groups of students across test administrations since 2009. The overall results for students in both the 4th and 8th grades have increased. However, only 38% of 4th graders and 34% of 8th graders met the proficiency benchmark in science. The scores for 12th graders on the assessment stayed flat from 2009, with only 22% of students scoring at or above proficiency. Despite the progress made among the 4th and 8th graders, all three groups remain well below the "proficient" level.

Group Differences

Gender. The fact that fewer females attain degrees in the STEM fields is in part attributed to differences in science achievement at the K-12 level (Sadler, Sonnert, Hazari, & Tai, 2007). Women remain underrepresented in the science and engineering workforce, although to a lesser degree than in the past, with the greatest disparities occurring in engineering, computer science, and the physical sciences (NSF, 2014). The results of a US census survey in 2012 indicate that only 1 in 7 women with a bachelor's degree in a STEM subject find work in a STEM field (U.S. Census Bureau, 2012). Previous studies using NAEP data have shown gender gaps in science achievement beginning as early as 3rd grade (e.g., Kohlhaas, Lin & Chu, 2010). However, evidence from newer research indicates that the beginnings of disengagement with science for girls begins earlier than middle school. More recently, studies using the ECLS data have revealed gender differences in science achievement starting in 1st grade (e.g., Curran & Kellogg, 2016).

When looking at gender differences in science achievement on the NAEP, at the 4th grade level, there was no significant difference between the average scores of girls and boys in 2015. For students in the 8th grade, boys scored on average 3 points higher than girls, down from a 5-point gap in 2011. There was not a significant difference among 12th graders between 2009 and 2015.

Although girls typically are outperformed by boys on state and national assessment in science like the NAEP, studies show girls and boys typically have similar results in science achievement in school (Else-Quest, Mineo, &Higgins, 2013). Science achievement differences by gender are thought to be associated less with lack of ability than with factors related to stereotype threat, cultural messaging, or self-efficacy (e.g., Good, Aronson, & Harder, 2007; Riegle-Crumb, King, Grodsky, & Muller, 2012). Student attitudes about science have been shown to be lower for elementary school girls than for boys (Pomerantz, Altermatt, & Saxon, 2002). A recent study has shown that starting at ages 6 and 7, girls begin to associate high level cognitive ability (the term used by the researchers is "brilliance") with males more often than with females (Bian, Leslie, & Cimpian, 2017). Patterns of thought influenced by stereotypical beliefs may persist into higher grades and influence academic and career trajectories. The intersection of gender and race/ethnicity is also an avenue for exploring differences in science achievement for girls. A study by Weinburgh (2003) of gender and attitudinal differences about science with African American students in fifth grade found that African American girls experienced overall less value of science to society and reduced motivation to engage in science than African American boys. However, Else-Quest, Mineo, and Higgins (2013), found no significant differences when examining the interaction between gender and race/ethnicity in a study of 10th grade students using measures of science achievement and attitude.

Race/Ethnicity. The racial/ethnic academic achievement gap is well-documented in the research literature. On state and national assessments for science, White students score higher on average than students from all other racial/ethnic groups, particularly when compared to Black and Hispanic students (NAEP, 2016). The gap between Black students and White students has been narrowing across the most recent test administrations of the NAEP among students in the 4th and 8th grades. The gap between 4th grade Black and White students decreased from 36 points in 2009 to 33 points in 2015. Similar results were seen for 8th grade students, with the divide in achievement between Black and White students narrowing from 36 points in 2009 to 34 points in 2015. At the 12th grade level, however, there was not a significant reduction in the gap. The gap between White students and Hispanic students in science on the NAEP has also decreased since 2009. The gap between White 4th grade students and Hispanic students narrowed from a 32-point difference on average in 2009 to 27 points in 2015. For 8th graders, the gap between White and Hispanic student scores went from a 30-point difference in 2009 to 26 points in 2015. The gap remained virtually unchanged at the 12th grade level. Despite these slight changes, the National Assessment of Educational Progress data has continued to indicate significant racial and ethnic achievement gaps that are predictive of inequitable outcomes for students after high school.

Outside of the results of national and international assessments, differences in achievement by race/ethnicity are found in other content areas, such as reading and math. Some of these differences indicate that, for some groups of students, significant differences exist even before students begin school (Fryer & Levitt, 2004; Phillips, Crouse, & Ralph, 1998). The ability to examine the beginnings of science achievement and growth for students of different race/ethnicities in this study shed light on the origin of and trajectory for these children.

Socioeconomic status. There is a wealth of information describing how socioeconomic status and the effects of poverty are important factors in explaining differences in educational achievement. Kindergarteners from families with incomes below the federal poverty levels scored lower on assessments than those from families with incomes higher than the federal poverty level, with assessment scores increasing with increasing parental education level (Mulligan et al., 2012). Overall, students from disadvantaged backgrounds continue to lag behind their more advantaged peers, with these disparities starting as early as kindergarten (NSB, 2014). The effects of neighborhood level poverty rate (socioeconomic segregation) and racial segregation have been demonstrated to have an effect on general academic achievement for students (e.g., Sastry & Pebly, 2010; Sharkey & Elwert, 2011), including such factors as disparity in access to school level resources and access to college preparatory classes in the sciences. Family SES has been shown to help to explain differences in STEM coursework participation and achievement in math and science (Miller & Kimmel, 2012). Students from families with low SES may have had unequal opportunities to be exposed to science content and knowledge, due to such factors as parent educational backgrounds, lack of resources allowing students to access science related resources at home or in informal or after school settings, and the availability of parent encouragement of science participation (Archer et al., 2012; Cotton & Wikelund, 2001; Harackiewicz, Rozek, Hulleman, & Hyde, 2012). Amplifying effects of SES on science achievement is the fact that students from racial/ethnic minority backgrounds are about twice as likely to live in poverty as those who are White (U.S. Census Bureau, 2012b).

Math and reading achievement. Previous studies have shown that science achievement is related to students' ability in math and reading (Snow, 2010). Most science classes require that students engage with more complex text containing discipline-specific vocabulary. Depending on the class, a wide range of ability with mathematical computation and analysis is often needed to demonstrate understanding of a given topic. Students with less well-developed reading and math skills may encounter difficulty with the learning progressions necessary to develop scientific reasoning and understanding. For some students, higher reading skills may help compensate for lower science background knowledge (O'Reilly & McNamara, 2007). More recent work by Reed, Petscher, and Truckenmiller (2016) indicates that intermediate reading skills may be more important to science achievement than previous studies have suggested. In their study, the authors found that general reading ability accounted for a greater percent of the variance in science scores for students in 5^{th} and 8^{th} grade on two standardized tests than hypothesized. Given that achievement gaps in math and reading have been shown to appear early in a student's academic trajectory and that these gaps are also often tied to race/ethnicity (e.g., Phillips, Krouse, & Ralph, 1998; Reardon & Galindo, 2009) and socioeconomic status (e.g., Morgan, Farkas, & Wu, 2009; Quinn, 2015), math and reading achievement are included in this study to help explain science achievement gaps.

Growth Trajectories

Morgan, Farkas, Hillemeier, and Maczuga (2016) suggested that it would be useful to examine achievement gaps between different groups of students in science from a developmental perspective, seeking to understand how science achievement trajectories may begin and possibly change over time by different grade level. Gaps in science achievement may begin to occur early due to lack of opportunities to interact with science learning (e.g., Archer et al., 2012; Sackes, Trundle, Bell, & O'Connell, 2011; Entwisle, Alexander, & Olson, 2000) then stay stable because science learning in school also stays stable across time and does not significantly alter the growth trajectories for any groups of students (McCoach, O'Connell, Reis, & Levitt, 2006).

Another growth trajectory possibility, often referred to as the Matthew effect (Stanovich, 1986), occurs when students enter school behind their peers in terms of science knowledge opportunities then experience low science achievement that continues to lag behind their peers over time (Morgan, Farkas, & Wu, 2011; Walberg & Tsai, 1983). In this model, also known as a cumulative advantage growth trajectory, differences in science achievement continues to grow for groups of students for a number of reasons, possibly including the increasing complexity of science instruction that occurs naturally as students progress to higher grades (Duncan, Rogat, &Yarden, 2009) and the fact that increased reading and mathematics skills are required in order to access more complex science content in the later grades (Fang, 2011; Snow, 2010).

Finally, there is a compensatory achievement growth trajectory, where students who initially lag actually catch up once they are exposed to science instruction (Huang, Moon, & Boren, 2014). In this model of science achievement trajectories, students may enter formal schooling with less science experience/lower achievement than their peers for the same reasons as mentioned before. However, their initial gaps may lessen over time as they experience schools with strong science programs, science interventions designed to help support students at risk, or informal science education in after-school programs (e.g., Museus et al., 2011; Wang, 2013; NRC, 2009). Students whose first language is not the language of instruction may see gains over time as well as their proficiency with the language of instruction increases (Crosnoe, 2012).

School Level Factors

Science education typically includes the disciplines of biology, chemistry, physics, and Earth and space sciences. Science as an instructional topic in the U.S. is a regular part of most K–12 students' school experiences. As with mathematics, specialized science classes begin in middle/junior high school. Testing in science under NCLB was mandated later (2007) than for mathematics and reading (2003) and at a much lower frequency. Unlike the testing for mathematics and reading, science was never part of the "adequate yearly progress" requirement that holds schools accountable for students' progress from year to year.

Science instruction in the elementary grades is typically, although not always, accomplished by the regular classroom teacher. Most elementary teachers are "generalists", meaning that the same teacher provides instruction in all of the content areas. Many educators do not feel confident teaching science, especially in the elementary and middle grades, and have lower content area knowledge, especially in the area of physical science (Appleton, 2006). According to the 2012 National Survey of Mathematics and Science Education, only 39% of elementary school teachers feel very well prepared to teach science, while 77% feel very well prepared to teach mathematics and 81% feel very well prepared to teach English language arts (Banilower, 2013).

According to the 2012 National Survey of Science and Math Education, just 5% of elementary teachers had a degree in science or science education, and 4% had a mathematics or mathematics education degree. Among middle school science teachers, 41% reported having earned a degree in science or science education, and 35% of middle school mathematics teachers had a degree in mathematics or mathematics education. The comparable figures for high school teachers were 82% and 73% for science and mathematics, respectively (Wenner, 2001). The nature of instruction in the STEM-related subjects requires a deeper understanding of the associated pedagogical content knowledge (PCK) for a particular area. Although teachers in the middle/secondary grades may have a firmer grasp on the STEM content knowledge than elementary school teachers, teachers at all grade levels would benefit from a better understanding of PCK, especially in conjunction with problem-solving and inquiry-based instructional practices (Loughran, Mulhall, & Berry, 2004). Professional development, used to increase teacher knowledge, skills, and self-efficacy (and thereby effectiveness), has changed through the years to become more focused and intentional, but the extent to which teachers have access to professional development in science is highly variable. In STEM subjects, high quality and targeted professional development is needed to provide teachers with a deeper understanding of the content and pedagogy specific to their discipline (Culp, Honey, &

Mandinach, 2005). Many of the available STEM-related curriculum materials do not do a good job in supporting problem-solving or inquiry-based science instruction. Additionally, schools may have curriculum materials that are not aligned to the NGSS and do not provide their teachers with sufficient training to adjust their current curriculum to align with the recently released NGSS, which may result in a misalignment with assessments designed to examine proficiency with NGSS standards. Educators need access to curriculum and supplemental materials that help to engage students in relevant exploration and connection with STEM-related topics and activities (Kesidou & Roseman, 2002). For these reasons, elementary school teachers not only often steer away from teaching science, but also use instructional strategies that may not be the best fit for the content area (WestEd, 2011). Over the past decade, there has been a trend in elementary schools toward spending less time on science and more time on reading and mathematics, presumably due at least in part to the NCLB legislation (Blank, 2012). Science is taught less frequently than mathematics, with only 20% of Grades K–3 classes and 35% of Grades 4–6 classes receiving science instruction all or most days, every week of the school year. Many elementary classes receive science instruction only a few days a week or during some weeks of the year. Elementary science instruction averages out to only 19 minutes on science per week. (Banilower, 2013).

At the school level, differences in access to high-quality science instruction and supports create opportunity gaps for some students more than others (e.g., Flores, 2007; Ladson-Billings, 2006). Students who attend schools in high poverty neighborhoods often have less access to high-quality instructional materials or lab facilities. Science classes in high-poverty schools are more likely than those in low-poverty schools to be taught by teachers with five or fewer years of experience (Peske & Haycock, 2006). While many science teachers in the secondary grades feel well prepared to encourage the participation of females and to encourage student interest overall in science and/or engineering, the proportion of teachers with this level of confidence decreases in the elementary grades and with teachers who have less teaching experience. At all grade levels, many teachers report that they do not feel well prepared to encourage the participation of students from low socioeconomic backgrounds and from racial or ethnic minorities in science and/or engineering. Additionally, few teachers indicate feeling very well prepared to teach science to students who have learning or physical disabilities, or are English-language learners (Banilower, 2013)

The national conversation about the lack of graduates ready to enter the STEM careers typically focuses on the number of students receiving bachelor’s degrees in STEM fields and the number of students who actually enter STEM careers. While the coursework for STEM degrees begins at the undergraduate level, the groundwork for STEM studies is laid during high school. Many courses in high school math and science act as gateway courses to continuing education pathways in STEM. High school courses required for the post-secondary STEM pipeline include advanced lab based biology, chemistry and physics coursework in addition to advanced math, typically calculus/pre-calculus. Earlier than high school, students on track to complete advanced math courses in high school need to have passed an algebra course while in middle school. Specific coursework in high school is required to lead to seamless transitions to STEM degrees and careers without entry-level remediation. Early remediation in math or science courses is related to the likelihood of attrition from STEM degrees (Miller & Kimmel, 2012)

Additionally, poor academic performance in the sciences in high school often has effects on the choice of a science-related major in college (e.g., Conley, 2007; Robinson & Ochs, 2008).

Previous Research on Science Achievement Gaps Using ECLS Data

At the time of the literature review conducted for this study, there were seven published, peer-reviewed articles that examined achievement in science using the ECLS dataset, with only one of the studies conducted using the ECLS-K: 2011 data. Of these studies, four focused on science achievement starting in Grade 3 or above (Kohlhaas, Lin, & Chu, 2010a; Kohlhaas, Lin, & Chu, 2010b; Lin, 2014; Quinn & Cooc, 2015). Chapin (2006) looked at the relation between children's social studies and science knowledge and skills to their race/ethnicity and gender for K-1 grade students using the ECLS-K data set. Morgan, Farkas, Hillemeier, and Maczuga (2016) examined science achievement gaps and growth trajectories across elementary and middle school in the ECLS-K study; however, the ECLS-K dataset did not have a separate science assessment in kindergarten and first grade, rather a General Knowledge assessment that included science items. A separate science assessment was not administered in the ECLS-K until the third-grade round of data collection. Most recently, Curran and Kellogg (2016) used the ECLS-K: 2011 dataset in a study on science achievement gaps by race/ethnicity and gender for students in grades K-1 and the contribution of individual student level factors and school fixed effects to explain identified achievement gaps. Findings from this study indicated significant science achievement gaps, particularly by race/ethnicity, and that the size of many of the detected achievement gaps change in between kindergarten and first grade.

While gender difference in science achievement was not significant in kindergarten, the emergence of a gender gap was noted in first grade.

Purpose of Study

The purpose of the present study was to build on the current interest in science achievement at the elementary level and leverage the availability of a nationally representative longitudinal sample that includes a measure of science achievement. The study was designed to investigate the science achievement growth trajectories of students by gender and race/ethnicity between kindergarten and Grade 2, using multilevel longitudinal models and applying rigorous statistical examinations of the results. This research adds to the literature around science achievement in several ways. First, the focus of attention is shifted away from the secondary and post-secondary years. As new research emerges that connects cognitive development in pre-school age children to the development of skills that are foundational to science understanding, it becomes clear that stakeholders and policymakers need to pay closer attention to the development of science achievement gaps and differential group trajectories if change is to be effected further down the educational pathway. Second, the results of this study extend the limited information currently available on achievement growth in science in the elementary school years and to call attention to the various student-level factors that may contribute to inequities in science achievement that persist into the middle and secondary years. These same factors, such as gender, race/ethnicity, and socioeconomic factors have been identified as having differential effects on student math and reading achievement. Observing the extent to which student level factors act in a similar fashion with science achievement helps to better inform intervention and instructional decision-making.

Finally, an instructional variable was included in the study variables to investigate if the amount of time spent on science instruction had a significant effect on student achievement and growth. Given that there is a high degree of variability in the amount of time spent on science instruction in the early school years, it could be surmised that students missing out on formal and explicit science instruction may be at a disadvantage to their peers in classrooms or schools that have a more coordinated or systematic approach to science instruction.

Research Questions

The research questions that were used to guide this study were as follows:

1. What is the average initial level and growth trajectory of students from kindergarten to second grade in science achievement?

2. To what extent does initial level of science achievement relate to growth in science achievement?

3. How are kindergarten to second grade growth trajectories affected by children's gender, race/ethnicity, family SES status, home language, and IEP status?

4. To what extent is kindergarten to second grade science achievement growth explained by student reading and math achievement in the fall of kindergarten, once demographic data are controlled for?

5. To what extent is kindergarten to second grade achievement growth explained by the average science instructional time, once demographic data and reading and math achievement are controlled for?

6. Is the effect of instructional time moderated by student demographics, specifically the interaction between girls and race/ethnicity?

CHAPTER III

METHODS

Study Data

Data used in this study were from the newest version of the Early Childhood Longitudinal Study, Kindergarten Class of 2010-11 (ECLS-K: 2011), conducted by the National Center for Educational Statistics (NCES). The ECLS-K: 2011 is a revised and updated version of the ECLS-K study, which began collection in 1998–1999. The 2011 study included a nationally representative sample of children who were either attending kindergarten or were 5 years old and attending school in an ungraded setting in the 2010–2011 school year. The base year sample consisted of 18,174 students who were selected to participate in the kindergarten year and were followed through fifth grade. The ECLS-K: 2011 employed a complex three-stage sampling strategy to create a nationally representative sample of public and private school students. In the first stage, the sampling design used stratified random sampling first within primary sampling units (PSUs) across the United States. Sampling of PSUs used 2007 census bureau population estimates for each county and estimates of the number of 5-year-old children in each (overall and by race/ethnicity), divided into regions (Northeast, Midwest, South, West). In order to sample PSUs with a probability proportional to size, each PSU was assigned a measure of size (MOS). The measure of size used for selecting the PSUs was the number of 5-year-old children in the PSU (rather than the total PSU population size) adjusted for the desired oversampling of Asian, Native Hawaiian, and other Pacific Islander students (designated APIs). In the next stage, approximately 1,300 schools were sampled from within PSUs. Finally, about 23 kindergarteners were sampled from within each selected school. If a sampled school had less than 23 kindergarten students, all of the students were selected for participation. The sample design included over-sampling Asian/Pacific Islander (API) students to ensure sufficient sample size, with API students sampled at a rate 2.5 times higher than non-API students (Tourangeau, Nord, Lê, Sorongon, et al., 2015; Tourangeau, Nord, Lê, Wallner-Allen, et al., 2017). The sampling design included an attrition without replacement strategy. Student replacement only occurred at the very beginning of the study when not enough schools in a region agreed to participate, but this was school replacement and not individual student replacement. Students selected to participate who did not actually participate in the base year activities were not contacted in later rounds of the study. According to the user's manual, 68% of base year students were still attending their original sampled school in the fall of second grade, 2013 (Tourangeau, Nord, Lê, Wallner-Allen, et al., 2017).

Participants & Procedures

The ECLS-K: 2011 study includes direct measures of both academic and cognitive skills and the physical measurement of student participants (including height and weight), as well as demographic and self-report survey data collected directly from children, their families, teachers, and schools (Tourangeau, Nord, Lê, Wallner-Allen, et al., 2015). Assessments and surveys were administered in the fall and the spring of kindergarten (2010-11), the fall and spring of first grade (2011-12), the fall and spring of second grade (2012-13), the spring of third grade (2014), the spring of fourth grade (2015), and the spring of fifth grade (2016). The sampling methodology used in the ECLS-K: 2011 included the use of fall *subsamples* in both the first and second grade years of data collection. The fall subsamples included children in approximately one-third of the sample of primary sampling units (PSUs) selected for the study. In the spring of first and second grade, the data collection included the entire sample of students participating in the base year collection, as did the single spring data collection time point in the third through fifth grade years. Data used in this study included publicly available data from students followed from kindergarten through the second-grade administration. See Table 1 for approximate sample sizes of participants ***for whom complete data*** was collected by data collection round (NCES, 2017).

Table 1

Approximate Sample Sizes (Rounded to the Nearest 10) of Student Participants in the Direct Child Assessments in the ECLS-K: 2011, K – 2

Data collection round	*n*
Fall kindergarten	13,760
Spring kindergarten	17,210
Fall first grade[1]	5,230
Spring first grade	15,130
Fall second grade[2]	4,740
Spring second grade	13,850

Note. [1]Assessments in the fall of first grade included a sub-sample of approximately one-third of the total ECLS-K: 2011 sample of students. [2] The assessments in the fall of second grade included the same students selected for the fall of first grade. Source: NCES ECLS-K: 2011 Training Handbook, 2017 Biennial Meeting of the Society for Research in Child Development. April 5, 2017. Austin, TX.

Assessment administration. The ECLS-K: 2011 assessments were individually administered to participating students by trained and certified child assessors. Assessors entered student responses using a computer- assisted interviewing (CAI) program. While the assessments were untimed, the total time involved for the direct cognitive assessment was designed to last for approximately 60 - 80 minutes, depending on the data collection round. The sequence of assessments delivered in the session for students in Grades K – 2 was as follows: Reading, Math, Executive Function, Science (except in the fall of kindergarten), and Height/Weight Measurements.

A language screener and routing path was used for all children taking the direct cognitive assessment, serving as a practice session for students whose primary language was English and as a screening mechanism for students whose primary language was something other than English. In kindergarten and first grade, adjustments were made to the assessment administration for students whose home language was one other than English or Spanish and who did not achieve at least the minimum score on the screener. Beginning in second grade, all study children were administered the entire assessment in English, regardless of home language.

An adaptive, two-stage approach to administration of the assessments was used for reading, math, and science. The first stage consisted of a routing section that included items across a range of difficulty. A student's performance on the routing section then determined which version of three second-stage assessments the student was administered, designed to include low, middle, or high difficulty items. The second-stage forms included some items that overlapped (e.g., some items in the low-level form also are included in the middle-level form). The common routing test and the item overlap between second-stage forms was used to ensure a sufficient number of items to precisely measure the child's skills. Specific cut scores were developed through statistical simulations that employed IRT ability estimates and item difficulty parameters in order for test takers to be assigned to a high, medium, or low-level assessment, as appropriate for their individual ability (Tourangeau et al., 2015). While most of the items from the first-grade assessments were included in the second-grade assessments, more difficult items were added to the assessments in second grade, and some easier items were removed. This process ensured that the assessments items were appropriate for the level of knowledge and skills of the students as they progressed through school (Tourangeau, Nord, Lê, Wallner-Allen, et al., 2017).

Across assessment domains, assessment items were presented to students on a small easel, with items typically consisting of images, words, number/number problems, depending on the domain. Some reading was required of students on the reading assessment, but all items in the math and science assessments were read aloud by the assessors. Students provided answers either verbally or by pointing at their selected responses on the test easel (Tourangeau, Nord, Lê, Wallner-Allen, et al., 2017).

ECLS-K Measures

The direct cognitive assessments used in the ECLS-K: 2011 study were created by expert curriculum specialists. Domains and constructs across the content areas of science, reading, and mathematics were identified and aligned with state and national standards. An item pool for each content area was developed using items from several published assessments, resulting in criterion-referenced, adaptive measures of cognitive ability. The ECLSK: 2011 user's manual presents expert review by curriculum specialists and examinations of performance standards as evidence for the validity of the direct cognitive assessments.

IRT-based scores. Item Response Theory (IRT) theta scores ranging from -6 to 6 were generated for each student, the higher scores indicated better performance on the assessment. Because academic achievement was reported using IRT scoring, student scores in the ECLS-K: 2011 data set represent an estimation of each student's ability (theta) in each domain at each measurement point. The resulting scores are reported on a linked scale so scores can be compared over time and across measurement occasions. Item response theory theta scores are used in the present study because they are comparable across different forms of an assessment and can be used to show growth over time (Tourangeau et al., 2017).

Reading achievement measure. The reading cognitive assessment included items to measure students' basic skills (phonological awareness, familiarity with print, recognition of letters and sounds, and identification of common sight words), vocabulary, and comprehension (locate/recall, integrate/interpret, critique/evaluate). There were 40 routing items on the first stage reading assessment in both the kindergarten and first grade years (Tourangeau et al., 2015) and 29 routing items in the second grade (Tourangeau et al., 2017). There was a total of 120 items in the reading assessment item pool across years. See Table 2 for the reported reliability of the IRT theta and scale scores (overall ability estimates) for the reading assessments for data collection conducted in grades K – 2 (Tourangeau et al., 2017).

Math achievement measure. The mathematics cognitive assessment included items to measure students' skills in conceptual knowledge, procedural knowledge, and problem solving, including number properties and operations, measurement, geometry, data analysis and probability, and algebra. There were 18 routing items on the first stage math assessment in both the kindergarten and first grade years (Tourangeau et al., 2015) and 20 routing items in the second grade (Tourangeau et al., 2017). There was a total of 113 items in the math assessment pool across years. See Table 2 for the reported reliability of the IRT theta and scale scores (overall ability estimates) for the math assessments for data collection conducted in Grades K – 2 (Tourangeau et al., 2017).

Science achievement measure. Unlike the original ECLS - K study, a separate direct measure of student science knowledge and skills was administered beginning in the spring of kindergarten. The science cognitive assessment included questions from the physical sciences, life sciences, environmental sciences, and scientific inquiry. For the spring kindergarten science assessment, the assessment was composed of 20 items that all children who were administered the science assessment received; a two-stage assessment was not used for this domain until the fall first-grade round. In first grade, there were 15 routing items on the first stage science assessment (Tourangeau et al., 2015) and 19 routing items in the second grade (Tourangeau et al., 2017). There was a total of 64 items in the science assessment item pool across years. See Table 2 for the reported reliability of the IRT theta scores (overall ability estimates) for the science assessments for data collection conducted in Grades K – 2 (Tourangeau et al., 2017).

Table 2

Reported Reliabilities of IRT Theta Scores for Reading, Math, and Science Achievement Assessments Used in the ECLS-K: 2011, by Data Collection Time Point.

Measure	K Fall	K Spring	Grade 1 Fall	Grade 1 Spring	Grade 2 Fall	Grade 2 Spring
Reading	0.95	0.95	0.95	0.93	0.93	0.91
Math	0.92	0.94	0.93	0.93	0.92	0.94
Science	---	0.75	0.83	0.83	0.83	0.83

Note. Science achievement measure was not administered in the fall of the baseline year. Source: *Early Childhood Longitudinal Study, Kindergarten Class of 2010–11 (ECLS-K:2011) User's Manual for the ECLS-K:2011 Kindergarten–Second Grade Data File and Electronic Codebook, Public Version* (Tourangeau et al., 2017).

Assessment psychometric data. Complete psychometric data are currently unavailable for the direct cognitive assessments used in the study. Additional documentation is currently in press (Najarian et al., forthcoming).

General classroom teacher questionnaires. Participating general classroom teachers completed self-administered paper teacher-level questionnaires in the fall and spring of each project year. The questionnaires included items about themselves and their classrooms, as well as questionnaires about each child in their classroom who was participating in the ECLS-K: 2011. Teacher-level questionnaires provided information about participating students' academic and social development and the classroom environment in which it occurred. Topics covered on the fall questionnaire included classroom and student characteristics and teacher demographic information. Topics covered on the spring questionnaire included instructional activities, content coverage, resources and materials, teacher evaluation and grading practices, and level of parent involvement.

Study Variables

Dependent variable. Analyses in this study used the IRT theta scores from the ECLS-K science assessments as outcome measures for science achievement growth. Science achievement was measured on 5 assessment occasions, beginning in the spring of kindergarten.

Demographic predictor variables. The demographic data used in this study to explain variation in student science achievement growth include sex, race/ethnicity, socioeconomic status, student disability status, and home language. These data were collected at the beginning of the study in the fall of 2010 through the use of school data systems, parent surveys, and student reports.

Student gender was dichotomously coded to represent male (0) and female (1). Race and ethnicity was coded as a nominal variable using the following categories based on the dataset: Black or African American (Non-Hispanic), Hispanic (race specified or non-specified), Asian, and Indigenous/Multiracial (including Native Hawaiian/Other Pacific Islander, American Indian or Alaska Native, More Than One Race). The small sample sizes of students in the categories included in Indigenous/Multiracial necessitated the collapsing of these groups into a single variable. White (Non- Hispanic) was used as the reference group for all analyses.

The variable representing socioeconomic status (SES) in this dataset was created as a composite using items collected in the parent survey including (a) male guardian's level of education, (b) female guardian's level of education, (c) male guardian's occupation, (d) female guardian's occupation, and (e) household income. Guardians' occupations were converted to prestige scores using the General Social Survey of 1989 (Tourangeau et al., 2015). Socioeconomic status is represented by a continuous variable on the scale of -3.00 to 3.00, with the larger values representing higher SES and the smaller values representing lower SES.

To control for possible differences in science achievement growth due to student disability status, a time invariant variable (designated IEP Ever) was created by combining data from spring teacher questionnaires. Each year, teachers were asked to indicate whether or not each student in their class who participated in the ECLS study had an Individualized Education Plan (IEP). Because of the small grade range of the present study, a decision was made to code a student as a 1 if they were reported as ever having an IEP at any occasion across the three reporting periods of the study. Students who were never reported as having an IEP were coded as 0.

The possible differences in science achievement growth due to student home language status were controlled for by including a dichotomously-coded variable that indicated whether or not a students' home language was reported as being English or another language. While an option exists on the teacher survey to indicate whether or not a participating student is identified as an English language learner, the author chose to use an indicator that represents a context-independent variable for students that may have a dominant language other than English, given possible validity issues with the EL variable that may arise as to whether or not a student is being treated as an EL student in school and differences in how a student is defined as an EL in one jurisdiction versus another.

Initial reading and math scores. As prior achievement measures, students' scores on the initial math and reading assessments from the baseline kindergarten assessment in the fall of 2010 were included in the model building process.

Science instructional time. Responses from items on the spring teacher questionnaire were used to create a student level variable regarding the average amount of time students received science instruction across the first three years of their schooling. The first relevant item on the questionnaire reads, "How often does the typical child in your class or classes usually work on lessons or projects in the following general subject areas, whether as a whole class, in small groups, or in individualized arrangements?" Response options for this item were: *Never*, *Less than once a week, 1 day a week, 2 days* *a week*, *3 days a week*, *4 days a week*, and *5 days a week*. The second relevant item on the questionnaire reads, "On the days children work in these areas, how much time does the typical child in your class or classes usually work on lessons or projects in the following general subject areas?" Response options included: *Not applicable/Never*, *Less than ½ hour a day*, *½ hour to less than 1 hour, 1 to less than ½ hours, 1 ½ to less than 2 hours*, *2 to less than 2 ½ hours*, *2 ½ to less than 3 hours,* and *3 hours or more*. The continuous instructional time variable was created by multiplying the midpoint of the selected time range by the number of days teachers reported science instruction. If the option of *Less than once per week* was selected, 0.5 was used as the multiplying factor.

Interaction effects. Stevens and Schulte (2016) demonstrated the importance of investigating interaction effects through the use of rigorous statistical analyses versus descriptive statistics or visual inspection as a means of avoiding misinterpretations in studies of achievement growth. In order to examine possible subgroup differences for girls belonging to historically underserved race/ethnicity categories related to the amount of science instructional time, interaction effect terms were created for Gender by Race/Ethnicity and were included in the final step of the model building process described in the next section.

Analytic Methods

The study dataset was created using the ECLS Electronic Codebook, which allows users to create a customized data extraction that only includes the variables of interest. Initial examinations of the data were conducted using SPSS 25 (IBM Corp., 2017). First, achievement scores and other variables of interest were examined for univariate normality, skew, and outliers using descriptive statistics, boxplots, and histograms. Descriptive statistics were used to report mean change by assessment time point for all students by student group. Next, student growth was modeled using multilevel longitudinal analyses (Raudenbush & Bryk, 2002). The multilevel analyses were completed using HLM 7.0 (Raudenbush, Bryk, Cheong, Congdon, & du Toit, 2011) and the longitudinal models were specified as random intercept and slope models. A third school level in the analysis was not included as the primary interest in this study was student growth rather than school effects. Thus, a two-level HLM model was used, with time centered at the final science testing occasion (spring Grade 2). An unconditional growth model served as a baseline for comparison to succeeding models. The next modeling step applied conditional multilevel growth models to examine model results when student demographic variables were added. The last step in this portion of the analysis was to add the remaining variables of interest (i.e., prior reading and math achievement, average time spent on science instruction across K - 2) as additional predictors of science achievement. At each step, nonsignificant terms were retained given that these substantive groupings account for the presence of variation that would be lost without their inclusion in the model. Differences between models were evaluated using deviance tests and calculation of pseudo-R^2 statistics. The conditional models included a level-1 model that specified student science scores predicted by a quadratic function of time of measurement and a level-2 model composed of the prediction of level-1 model parameters as a function of student demographic characteristics and time spent on science instruction. Individual growth curves were evaluated after initial examination of the data to fully inform the decision of what functional form of the science achievement growth trajectories to use. After specifying growth models in HLM, tests of homogeneity of

Level 1 variance were conducted (Raudenbush & Bryk, 2002). Residuals for all levels of the HLM models were examined in SPSS for consistency with assumptions underlying multilevel models. Residual files were used to create a scatterplot of the Mahalanobis distances against the expected chi-square distribution values from the higher-level residual files, including a fit line to ensure consistency with assumptions underlying multilevel models.

Level 1 Model:

$$Y_{ti} = \pi_{0i} + \pi_{1i}(Time_{ti}) + \pi_{2i}(Time_{ti}^2) + e_{ti} \tag{1}$$

where

Y_{ti} is the science IRT theta scale score for student *i* at time *t;*

π_{0i} is the initial status or intercept for student *i* at time 0;

π_{1i} is the linear rate of change;

π_{2i} is the quadratic curvature;

e_{ti} is the residual for student *i* at time *t.*

Level 2 Model:

$$\pi_{0i} = \beta_{00} + \Sigma\beta_{0k}(Predictor_k) + r_{0i} \tag{2}$$

$$\pi_{1i} = \beta_{10} + \Sigma\beta_{1k}(Predictor_k) + r_{1i} \tag{3}$$

$$\pi_{2i} = \beta_{20} + \Sigma\beta_{2k}(Predictor_k) + r_{2i} \tag{4}$$

where

β_{00} is the science IRT theta scale score intercept for student *i* at time *t* in the spring of second grade;

β_{0k} is the average partial regression coefficient relating the predictor of interest to student initial status;

r_{0i} is the residual between the fitted predictor value for student i and that student's observed science theta score;

β_{10} is the average rate of linear change per occasion;

β_{1k} is the average partial regression coefficient relating the predictor of interest to student rate of linear change;

r_{1i} is the residual between the fitted predictor value for each student's linear rate of change and the observed rate of linear change;

β_{20} is the average rate of quadratic change per occasion;

β_{2k} is the average partial regression coefficient relating the predictor of interest to student rate of quadratic change;

r_{2i} is the residual between the fitted predictor value for each student's quadratic rate of change and the observed rate of quadratic change.

Effect size. In order to empirically describe the magnitude of group differences and to provide assistance in interpreting the growth trajectories, science achievement growth effect sizes were calculated. The use of a longitudinal design necessitated the use of a growth effect size, which takes into account the parameters of the estimated growth model and intra-individual dependency between occasions (Stevens, 2018). The growth model effect size (ES) used here also takes into account non-linear growth and the inclusion of a quadratic term in this study's analyses. Using HLM residual files, Empirical Bayes estimates of growth model parameters were used to calculate individual growth trajectories. An estimated growth rate at each assessment time point was calculated, and serves as the metric of the magnitude of change. Growth rate (GR) was determined using the following formula: GR = S + (2)(Q)(t), where S and Q are the estimated linear slope and quadratic growth parameters in the growth model and t represents time or measurement occasion. In order to obtain an *annual* growth effect size, the growth rate was multiplied by 12 (because time was coded in months). To determine the growth ES, the growth rate was then divided by the standard deviation for all students at the grade level of interest (time t).

Sample weights. To ensure that data were collected from a wide range of children, a strategic sampling method was used in the ECLS-K: 2011 study to include a representative sample of schools, teachers, and children in participation, meaning that subjects did not have an equal chance of being selected for the study. Given the stratified, multistage probability sampling design, a longitudinal child level sample weight was applied to all analyses. The purpose of this sample weight was to account for differential probabilities of selection and nonresponse bias. Applying the weighting variable ensured that the analytic sample reflected a nationally representative sample to the degree possible using this method. The ECLS Electronic Codebook contains select weighting schemes that are provided for analysis using data from each round of data collection and components (i.e., assessments and interview questionnaires) for which the weight is adjusted for nonresponse. Appropriate weights for analysis were used, specifically the kindergarten through second-grade dataset weighting scheme that accounts for student assessment data at all time points and teacher survey data from the spring of each year (W6CS6P_2T0). To account for the possibility of bias given the sampling design and large sample size, design effect calculations were applied to the sampling weights to adjust the standard errors. The same sample weight was used for all analyses so that the underlying sample used to produce estimates remained the same (Tourangeau, Nord, Lê, Wallner-Allen, et al., 2017).

Missing data. The ECLS-K: 2011 study utilized a purposeful, stratified sampling plan and the dataset includes sampling weights that reflect this plan and account for sampling bias. Additional analyses of the dataset were conducted to determine if missingness was random or systematic. For variables with missing data greater than 2%, *t*-tests or chi-square tests were conducted to examine differences in each outcome variable based on missingness. The decision was made to use full information maximum likelihood estimation (FIML) at Level 1 and multiple imputation at Level 2 in the multilevel modeling process to deal with the missing data. FIML is preferred where possible, but, in this analysis, was only possible for Level 1, thus multiple imputation was used for Level 2. These methods were selected because of their abilities to more effectively and accurately improve the analyses relative to other methods of dealing with missing data (Enders, 2010).

CHAPTER IV

RESULTS

In this section, results from all of the relevant analyses are presented and discussed, along with visual displays to help in their interpretation.

Testing Assumptions

The science achievement IRT scale score was examined to determine if the distributions across assessment waves were normally distributed (see Appendix for distributions and scatterplots across all assessment time points). Results revealed relatively normal distributions without severe kurtosis (i.e., kurtosis statistics ranged from -0.29 to 2.67). However, while the first three assessment time points were without severe skew (i.e., skewness statistics ranged from -0.52 to 0.02), the distributions of science achievement scores for the final two assessment time periods were both slightly negatively skewed to the right (i.e., skewness statistics of -1.01 and -1.03, respectively).

The computed average instructional time variable displayed slight positive skew (skew statistic = 1.33) and a leptokurtic distribution (kurtosis statistic = 3.21), with a very wide range of reported instructional time, from 2.5 hours to 575 hours of average instructional time across the K-2 grade span (M = 126.58, SD = 69.99). Pearson's correlations revealed statistically significant correlations between many of the study variables (see Table 3).

Table 3

Pairwise Correlations Among Study Variables

Variable	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
1. Sex	1															
2. Black	.006	1														
3. Hispanic	.001	**-.222**	1													
4. Asian	.025	-.086	**-.122**	1												
5. Indigenous/ Multiracial	.000	**-.097**	**-.137**	**-.053**	1											
6. Home language	-.009	**.108**	**-.421**	**-.234**	**.069**	1										
7. SES	.004	**-.136**	**-.317**	**.072**	.017	**.265**	1									
8. IEP Ever	.010	.003	.016	-.006	.006	-.002	-.004	1								
9. Science SPK	-0.15	**-.181**	**-.321**	**-.072**	**.048**	**.345**	**.423**	-.011	1							
10. Science F1	-0.11	**-.148**	**-.335**	-.024	**.071**	**.322**	**.439**	-.003	**.804**	1						
11. Science SP1	**-.040**	**-.201**	**-.270**	-.026	**.051**	**.269**	**.421**	-.002	**.765**	**.840**	1					
12. Science F2	-.037	-.105	**-.305**	.012	.049	**.283**	**.430**	-.002	**.737**	**.797**	**.833**	1				
13. Science SP2	**-.046**	**-.210**	**-.229**	.030	.031	**.202**	**.397**	-.006	**.719**	**.777**	**.815**	**.856**	1			
14. Reading FK	**.047**	**-.062**	**-.203**	**.072**	.008	**.161**	**.400**	.000	**.515**	**.569**	**.549**	**.555**	**.541**	1		
15. Math FK	-.018	**-.132**	**-.240**	**.068**	.013	**.188**	**.410**	-.009	**.577**	**.606**	**.607**	**.595**	**.614**	**.762**	1	
16. Instructional time	.006	**.137**	**.070**	-.011	**-.054**	-.025	-.018	.007	-.038	-.038	-.032	-.033	-.015	.035	.008	1

Note. Bolding indicates statistical significance. Correlations calculated using pairwise deletion.

A visual inspection of a random selection of 25 samples of science achievement performance across Grades K–2 (see Figure 1) suggested the presence of potential quadratic form, resulting in the inclusion of a quadratic term in the model building process to account for individuals with nonlinear growth trajectories.

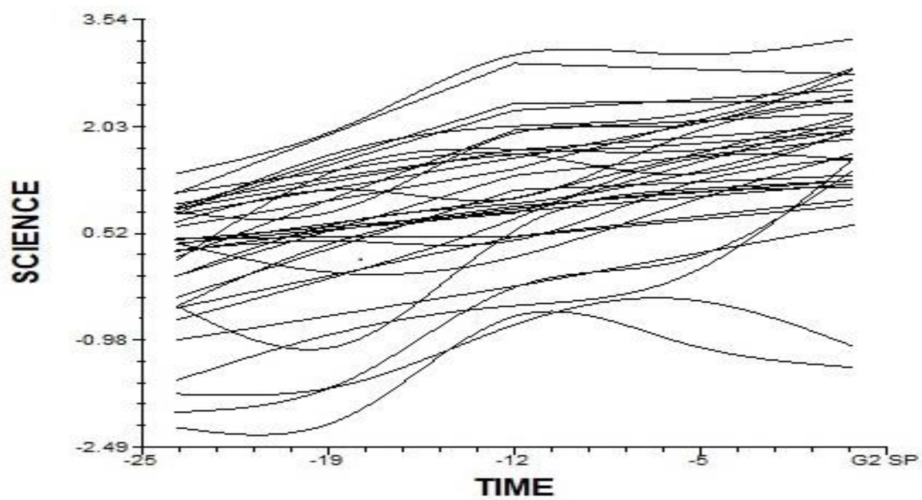


*Figure 1***.** Random sample of 25 science achievement growth trajectories

Missingness Analysis

The raw data set were first examined to create an analytic sample. The full sample size was 18,174 cases. Data on whether or not a student was not assessed at any time point due to a disability exclusion was suppressed in the public dataset, so this could not be determined in the present study.

In order to establish a baseline for the succeeding growth models, students who did not have an initial kindergarten science test score (n = 1,238) were excluded, reducing the sample size by 6.81% to 16,936. Students who were missing data from the initial kindergarten science assessment time point had statistically significantly lower socioeconomic status scores than students not missing the initial science theta score (t =

10.34, $p < .001$), as well as statistically significantly lower initial reading ($t = 13.74$, $p < .001$) and initial math scores ($t = 14.27$, $p < .001$). Students missing a baseline science assessment score were statistically more likely to be Hispanic ($\chi^2 = 41.53$, $p <.001$), or Asian ($\chi^2 = 12.01$, $p = .001$).

After this exclusion rule was applied, the percentages of missing data for the other variables of interest from the non-weighted sample were calculated and displayed in Table 4. Although Enders (2010) identifies several problems with the overall utility of Little's MCAR test, SPSS was used to conduct the omnibus test, resulting in a chi-square = 3,098.55 (df = 1,424, $p < .05$), indicating that, as a whole, the data in the sample were not missing completely at random.

Table 4

Number and Percent of Missing Data by Variable of Interest

Variable	Number of missing cases	Percent of sample
Sex	29	0.17%
Race/ethnicity	40	0.24%
Family SES	1,841	10.9%
IEP Ever	317	1.87%
Home language	1,881	11.1%
Kindergarten reading score	1,964	11.6%
Kindergarten math score	1,964	11.6%
Grade 1 spring science score	1,918	11.3%
Grade 2 spring science score	3,016	17.8%
Science instructional time	681	4.02%

Note. Missing data presented here from first analytic sample, after baseline exclusion rule applied, N = 16,936.

Students missing data for the socioeconomic status variable had statistically significant lower science achievement scores in the spring of kindergarten, Grade 1, and Grade 2 (t = 16.45, p < .001; t = 14.58, p < .001; t = 12.83, p < .001). Hispanic (χ^2 = 48.69, p <.001), Black (χ^2 = 44.08, p <.001), and Asian (χ^2 = 32.62, p <.001) students were also more likely to be missing data for the socioeconomic status variable than White or Indigenous/Multiracial students.

Students missing a response to the primary home language variable were statistically more likely to be male (χ^2 = 4.27, p <.05), black (χ^2 = 41.11, p <.001), Hispanic (χ^2 = 53.51, p <.001), or Asian (χ^2 = 32.69, p <.001). Students missing a response to the primary home language variable also had statistically significantly lower socioeconomic status than students with data for this variable (t = 2.86, p < .01).

While there were no statistically significant differences in socioeconomic status between students not missing and missing science scores in the spring of kindergarten or Grade 1, students missing science test scores in the spring of Grade 2 had statistically significant lower socioeconomic status (t = 3.51, p <.001). Students missing science scores in the spring of Grade 1 were more likely to have a home language other than English (χ^2 = 70.75, p <.001), more likely to be black (χ^2 = 92.27, p <.001) and more likely to be Hispanic (χ^2 = 40.81, p <.001). A similar pattern was identified for students missing science scores in the spring of Grade 2. Students missing scores at this time point were more likely to have a home language other than English (χ^2 = 79.03, p <.001), more likely to be Black (χ^2 = 130.60, p <.001) and more likely to be Hispanic (χ^2 = 65.69, p <.001), with the addition of Indigenous/Multiracial students being more likely to be missing science scores in the spring of Grade 2 (χ^2 = 21.21, p <.001).

Students missing reading scores in the fall of kindergarten had statistically significant lower science scores in the spring of kindergarten (t = 3.22, p < .01), in the spring of Grade 1 (t = 4.99, p < .001), and in the spring of Grade 2 (t = 2.91, p < .01). Students missing reading scores in the fall of kindergarten were also more likely to have a home language that was not English (χ^2 = 128.23, p <.001) and to be Hispanic (χ^2 = 39.08, p <.001) or Asian (χ^2 = 96.39, p <.001).

Similarly, students missing math scores in the fall of kindergarten had statistically significant lower science scores in the spring of kindergarten (t = 3.22, p < .01), in the spring of Grade 1 (t = 4.99, p < .001), and in the spring of Grade 2 (t = 2.91, p < .01). Similar to the missingness for reading scores, students missing math scores in the fall of kindergarten were more likely to have a home language that was not English (χ^2 = 128.23, p <.001), and to be Hispanic (χ^2 = 39.08, p <.001) or Asian (χ^2 = 96.39, p <.001).

Planned missingness. The fall administration of the cognitive assessments in Grades 1 and 2 included a planned missingness sampling design to reduce the overall burden of testing in the ECLS-K:2011 project. During each fall assessment period, one-third of study students were selected to participate in the cognitive assessments. After applying the previously described exclusion rule, missingness analyses found that 66.9% of students (n = 11,328) were excluded from the fall subsample across both years, while 33.1% of students (n = 5,608) were selected to participate in the fall subsample test administration. In Grade 1, 2.8% of students (n = 466) who were selected for fall testing were missing science achievement scores, while in Grade 2, 5.6% of selected students (n = 942) were missing science achievement scores.

For both the Grade 1 and Grade 2 subsamples, there were no statistically significant differences on SES or initial reading or math score in the fall of kindergarten between students who were in the fall subsample and did not participate versus those who did participate. Of the 16,936 students in the initial analytic sample, 25.6% (*n* = 4,344) had science achievement scores in all 5 test waves, 1.92% (*n* = 306) had scores in 4 waves, 52.9% (*n* = 8,963) had scores in 3 waves, 7.54% (*n* = 1,277) had scores in 2 waves, and 12.8% (*n* = 2,046) had only one science score during the first wave of assessment administration.

It is important to note that designs such as this that include planned missingness produce missing completely at random data, where the missingness of scores from students not selected to participate in the fall Grade 1 and 2 assessments is not related to the study variables. Given the complexity of the missing data patterns in the ECLS-K dataset, the degree of missingness constitutes a limitation to the validity of the results of the study analyses.

Multiple Imputation

Prior to conducting statistical analyses, multiple imputation was used to produce plausible predicted values for missing information in the Level 2 data file while preserving variability (Enders, 2010). Five imputed datasets were generated in SPSS v. 25 (IBM Corp., 2017). While some researchers (e.g., Graham, Olchowski, & Gilreath, 2007) would suggest that larger numbers of datasets should be imputed to improve the power of the analysis, given the large sample size, the author chose to follow Rubin's (1987) suggestion that five imputations would be sufficient. All student-level covariates used in the regression models were used in the imputation models.

After creating the complete data sets, each file was separately uploaded to HLM for estimating the growth models. HLM automatically combined the parameter estimates and standard errors into a single set of calculated results. However, because the average output file for each model specification run contained calculated results, no pooled deviance statistics are produced (M. du Toit, personal communication, March 30, 2018). A review of the literature on the appropriateness of pooling goodness-of-fit statistics suggested that this approach would not be acceptable and that there is not currently consensus on the best alternative to take in this case (see White, Royston & Wood, 2011; Enders, 2010). Accordingly, the author ran separate analyses on each individual data file, including the original Level 2 file with no imputed data, with the specific intent to review and compare the deviance statistics for each one. The overall results in terms of significance and goodness-of-fit for each model was the same across repeated analyses. Deviance statistics reported with the regression model results are from each of the five imputed files.

Design Effects

Because the Early Childhood Longitudinal Study, Kindergarten Class of 2010–11, data were collected using a stratified multistage random sampling design rather than simple random sampling techniques, descriptive statistics and associated standard errors were estimated taking the complex design into account. Failure to use appropriate adjustments could lead to the underestimation of the standard errors, biased parameter estimates, and poor performance of test statistics and confidence intervals (Hahs-Vaugn, 2005). The method selected for this study was to use an approximation method that involved two steps, leading to the creation of a new analytic weight variable.

First, the sample weights were normalized to ensure the standard error was based on actual sample size rather than population size. Next, the normalized weights were divided by the appropriate design effect (DEFF) provided in the ECLS-K: 2011 user's guide to account for the sampling design (Tourangeau, Nord, Lê, Wallner-Allen, et al., 2017). Analyses performed in both SPSS and HLM allowed for the application of the adjusted weights.

Analytic Sample

The application of the sample weights reduced the analytic sample to a total of 11,081 students with non-zero sample weights. While this reduction constitutes a loss of sample size compared to the original full sample, other researchers using the ECLS-K and ECLS-K: 2011 datasets have demonstrated stable findings through robustness checks and sensitivity analyses (e.g., Curran & Kellogg, 2016) and the reduction results in a representative sample that can be generalized to the population of interest.

Descriptive Statistics

Table 5 displays descriptive statistics for the weighted full analytic sample, disaggregated by student race/ethnicity and gender. The sample is fairly evenly divided between male and female students. It can also be seen that the sample is composed of an even split between White students and students from all other race/ethnicity categories. A high percentage of students who come from families where the primary language is not English were Hispanic (47.1%) and Asian (59.8%). White and Asian students have higher initial math and reading scores in kindergarten than their peers in other race/ethnicity categories, as well as coming from families with higher average socioeconomic status. Interestingly, Black students had the highest reported average science instructional time, followed by Hispanic and Asian students. However, the degree of variability associated with the science instructional time was also larger for non-White students.

Table 5

Weighted Descriptive Statistics for Key Predictor Variables

	Asian	Black	Hispanic	Indigenous /Multiracial	White	Female	Male
Variable	*n* = 122	*n* = 367	*n* = 651	*n* = 153	*n* = 1423	*n* = 1320	*n* = 1396
Female	54.1%	49.3%	48.7%	49.0%	47.9%	48.6%	51.4%
IEP Ever	13.1%	12.5%	12.3%	15.7%	12.4%	12.5%	12.8%
Home language other than English	59.8%	4.36%	47.1%	3.95%	1.97%	16.3%	15.3%
Fall K reading score	-0.26 (0.09)	-0.65 (0.04)	-0.81 (0.03)	-0.48 (0.07)	-0.36 (0.02)	-0.48 (0.02)	-0.55 (0.02)

	Asian	Black	Hispanic	Indigenous /Multiracial	White	Female	Male
Fall K math score	-0.20	-0.77	-0.85	-0.41	-0.24	-0.48	-0.44
	(0.08)	(0.04)	(0.04)	(0.07)	(0.02)	(0.02)	(0.02)
Composite SES measure	0.17	-0.35	-0.52	-0.03	0.16	-0.08	-0.09
	(0.08)	(0.03)	(0.03)	(0.06)	(0.02)	(0.02)	(0.02)
Average science instructional time	124.05	150.58	135.92	110.30	118.24	127.23	126.18
	(5.67)	(4.10)	(3.00)	(4.75)	(1.67)	(1.94)	(1.84)

Note. Standard errors in parentheses. Estimates are adjusted to account for the complex sampling design of the Early Childhood Longitudinal Study, Kindergarten Class of 2010–2011 (ELCS-K:2011) using sample weights and design effect estimates of the standard errors.

Table 6 displays the weighted means and standard deviations of science achievement theta scores, disaggregated by assessment time point and by each student group of interest. An examination of these scores indicate differences between groups in science achievement within and across Grades K – 2. As a group, White students have the highest incoming science scores in kindergarten, and this trend continues across all assessment time points. Black and Hispanic students have the lowest incoming science scores, and while Black students slightly overtake Hispanic students in the fall science assessment, the across grade trends remain the same. As a group, Asian students have incoming science scores closer to Black and Hispanic students, but across the five assessment time points, their science scores increase at a faster rate and end up higher than students from all other race/ethnicity categories except White students. When looking at gender differences, males and females across all race/ethnicity categories have similar incoming science scores, but scores for female student remain lower across the five assessment time points and start to widen over time. Both Black and Hispanic females have the lowest science scores across the five assessment time points.

Table 6

Weighted Science Theta Score Means and Standard Deviations by Student Group Across Grades

Student group	Kindergarten Spring	Grade 1 Fall	Grade 1 Spring	Grade 2 Fall	Grade 2 Spring
All students	0.04 (0.88)	0.35 (0.96)	0.87 (0.98)	1.15 (0.97)	1.61 (0.89)
Race/ethnicity					
Asian	-0.25 (0.89)	0.24 (0.98)	0.87 (0.96)	1.21 (1.03)	1.78 (0.89)
Black	-0.37 (0.84)	-0.03 (0.85)	0.55 (0.85)	0.87 (0.90)	1.28 (0.86)
Hispanic	-0.47 (0.90)	-0.08 (0.91)	0.41 (0.96)	0.75 (1.05)	1.26 (0.95)
Indigenous /Multiracial	0.21 (0.81)	0.61 (0.93)	1.10 (0.91)	1.33 (0.94)	1.73 (0.85)
White	0.38 (.70)	0.80 (0.81)	1.29 (0.83)	1.53 (0.74)	1.95 (0.71)
Gender					
All females	-0.07 (0.87)	0.34 (0.95)	0.84 (0.96)	1.11 (0.94)	1.57 (0.87)
All males	-0.04 (0.93)	0.37 (0.97)	0.87 (0.99)	1.18 (1.00)	1.64 (0.91)
Asian females	-0.18 (0.90)	0.26 (0.97)	0.89 (0.93)	1.21 (0.99)	1.79 (0.81)
Asian males	-0.24 (0.98)	0.22 (1.02)	0.84 (1.03)	1.20 (1.11)	1.78 (1.01)
Black females	-0.33 (0.83)	-0.04 (0.86)	0.58 (0.85)	0.85 (0.90)	1.32 (0.83)

Student group	Kindergarten Spring	Grade 1 Fall	Grade 1 Spring	Grade 2 Fall	Grade 2 Spring
Black males	-.32	-0.01	0.52	0.88	1.25
	(0.85)	(0.84)	(0.86)	(0.91)	(0.89)
Hispanic females	-0.53	-0.08	0.40	0.72	1.23
	(0.86)	(0.90)	(0.94)	(0.99)	(0.91)
Hispanic males	-0.54	-0.08	0.43	0.78	1.28
	(0.91)	(0.91)	(0.98)	(1.10)	(0.98)
Indigenous /Multiracial females	0.17	0.53	1.02	1.29	1.69
	(0.79)	(0.94)	(0.92)	(0.94)	(0.83)
Indigenous /Multiracial males	0.28	0.67	1.16	1.35	1.76
	(0.88)	(0.93)	(0.91)	(0.95)	(0.88)
White females	0.37	0.78	1.26	1.48	1.89
	(0.70)	(0.79)	(0.83)	(0.72)	(0.73)
White males	0.41	0.81	1.32	1.57	2.00
	(0.71)	(0.83)	(0.82)	(0.75)	(0.73)

Note. Standard errors in parentheses. Estimates are adjusted to account for the complex sampling design of the Early Childhood Longitudinal Study, Kindergarten Class of 2010–2011 (ELCS-K:2011) using sample weights and design effect estimates of the standard errors.

The range of science achievement theta scores by grade show that over time, the distribution of scores widened, indicating an increased variability in achievement between students starting at Grade 2. In kindergarten and Grade 1, the range of science theta scores remained fairly stable across assessment time points, at -2.41 – 3.59 in the spring of kindergarten, -2.34 – 3.59 in the fall of Grade 1, and -2.34 – 3.89 in the spring of Grade 1. However, in the fall of Grade 2, the range of science theta scores increased to -4.56 – 3.36, with an even wider range in the spring of Grade 2, at -4.47 – 5.36.

Model Estimation

A two-level model was estimated with Level-1 representing measurement occasions and the second level representing students and their characteristics. Robust standard errors were used to compensate for potential non-normality (Liang & Zeger, 1986). Estimation settings included the use of full maximum likelihood estimation, inclusion of the five imputed data sets, and the application of the normalized sampling weights at both Level 1 and Level 2. All dichotomous variables were uncentered, while initial math and reading achievement, socioeconomic status, and science instructional time, all continuous variables, were grand mean centered.

In the two-level HLM models used for analyses, time was centered at the final testing occasion in the spring of Grade 2. Time in months was coded across the five measurement occasions, as -24, -18, -12, -6, and 0. The time coding is potentially problematic given the highly irregular assessment administration schedule in the ECLS-K:2011 data set and are discussed further in the limitations section of the Discussion. For example, in the baseline year of the study, fall assessments occurred from August through December, while spring assessments occurred from March through June. Students did not necessarily have an equal spacing between assessment time points. In order to standardize the coding scheme for this study, the author examined the distributions of student assessments across each year of the ECLS study and used the months with the highest frequency of assessment in each testing window as anchors. This decision rule produced the six-month time coding scheme used in this study.

After specifying growth models in HLM, a test of homogeneity of Level 1 variance (Raudenbush & Bryk, 2002) indicated normally distributed errors across science achievement time points χ^2 (df = 10,278) = 2,471.20, p > .500. Residuals for all levels of the HLM models were examined in SPSS for consistency with assumptions underlying multilevel models. A Q-Q plot of the level 1 residuals for science achievement was fairly linear, however some outliers may indicate a degree of departure from linearity. Additionally, a visual inspection of a scatterplot of the Mahalanobis distances against the expected chi-square distribution values also indicate that outliers in the dataset may be a threat to the assumption of normality underlying multilevel modeling. This threat may pose potential limitations to the validity of the results of the analyses.

Unconditional models. The first step in the model building process was to apply unconditional growth models to serve as a baseline for comparison to succeeding models. The fully unconditional means model, which estimated grand means and variance components, resulted in an intraclass correlation coefficient (ICC) of .43, representing the proportion of the variance in outcome between the students. In this case, 43% of the variance is thus over time, and the remaining 57% at a student level.

Table 7 displays the random effects for the unconditional model using alternate time coding schemes. Biesanz, Deeb-Sossa, Papadakis, Bollen, and Curran (2004) describe the importance of understanding and interpreting the role of coding time in growth modeling. The first column shows the variance component, residual, and deviance statistics for all five imputed files when time is centered on the spring of kindergarten assessment time point (0, 6, 12, 18, 24). The linear slope varied significant across individuals (χ^2 = 12,608.16, p < .001), as did the quadratic slope (χ^2 = 12,827.38, p < .001) suggesting differing rates of linear and quadratic change among students at the initial administration of the science assessment. The second column shows the results when time is centered at the midpoint, the spring of Grade 1 assessment time point (-12, -6, 0, 6, 12). In this case, the model would not converge with the inclusion of a quadratic term. However, excluding the quadratic term resulted in a linear slope that varied significantly across individuals (χ^2 = 16,183.71, $p < .001$), suggesting differing rates of linear change among students at the midpoint of the assessment time period, but no overall quadratic curvature of the slope. The third column shows the results that were used in this study, with time centered on the final assessment time point in the spring of Grade 2. The linear slope varied significant across individuals (χ^2 = 12,195.03, $p < .001$), as did the quadratic slope (χ^2 = 12,827.01, $p < .001$). This final time coding was used because the research questions guiding the present study focused on growth at the second-grade time point. While some researchers have pointed out that centering time on the midpoint may be a desirable methodological choice given that this centering scheme reduces collinearity between the linear and quadratic time variables, the outcome is that the intercept is then interpreted as the value of the dependent variable at the middle time point rather than at the point of interest. Additionally, Biesanz and colleagues (2004) suggest that concerns about multicollinearity should be of less interest than an intentional focus on the questions of interest guiding the analysis. It is important to note in this section that because of the time coding scheme used in the following analyses, with time centered on Grade 2 and negative time codes extending back to the spring of kindergarten, the interpretation of the parameter estimates for the linear and quadratic effects should include consideration of the sign for that particular effect and its relation to the negative

Table 7

Random Effects for Unconditional Longitudinal HLM Regression Models for Alternate Time Coding Schemes, Grades K – 2

	Time Centered on SP K			Time Centered on SP 1			Time Centered on SP 2		
Random effect	Intercept	Linear	Quadratic	Intercept	Linear	Quadratic	Intercept	Linear	Quadratic
Variance component	**0.64**	**0.0002**	**0.001**	**0.66**	**0.0002**	---	**0.64**	**0.0006**	**0.00000**
Residual	0.14			0.16			0.14		
Deviance, File1			868.48			40,319			868.48
			(4, <.001)			(4, <.001)			(4, <.001)
Deviance, File2			868.48			40,319			868.48
			(4, <.001)			(4, <.001)			(4, <.001)
Deviance, File3			868.48			40,319			868.48
			(4, <.001)			(4, <.001)			(4, <.001)
Deviance, File4			868.48			40,319			868.48
			(4, <.001)			(4, <.001)			(4, <.001)
Deviance, File5			868.48			40,319			868.49
			(4, <.001)			(4, <.001)			(4, <.001)

Note. Estimates are adjusted to account for the complex sampling design of the Early Childhood Longitudinal Study, Kindergarten Class of 2010–2011 (ELCS-K:2011) using sample weights and design effects. Bolding indicates statistical significance.

sign for the time codes. Thus, the linear term describes growth over a short time period with the *opposite* effect of the sign ascribed to a parameter estimate. For a quadratic term, the sign should be interpreted as usual, a positive parameter estimate indicating acceleration while a negative parameter estimate indicating deceleration.

Next, a two-level model that included a quadratic term and used the spring of Grade 2 coding scheme was fully examined (see Table 8, column 1 for results). The inclusion of the quadratic term resulted in a statistically significant improvement in model fit ($p <. 001$). For this model, the estimated mean science theta score across students at the spring Grade 2 time point was 1.62. Both the average linear change and the quadratic change were statistically significant. The linear change was 0.06 ($t = 64.73$, $SE = 0.001$, $p < .001$), while the quadratic component indicated a positive curvilinear trend at 0.0003 ($t = 10.46$, $SE = 0.00004$, $p <.001$). Intercorrelations of the model parameters were -0.58 (intercept and linear), 0.64 (intercept and quadratic), and -0.82 (linear and quadratic).

Conditional models. The next step in the model building process was to estimate three conditional models. The first conditional model included the addition of student level predictors that were hypothesized to have differential effects on student science achievement growth. The second column in Table 8 shows the results when student gender, race/ethnicity, home language, disability status, and family socioeconomic status were entered. At this point, the intercept represents the mean science theta score in the spring of Grade 2 for White males whose home language is English and do not have an IEP across the Grade K – 2 time span ($M = 1.86$, $SE = 0.01$). The results show that female students had a statistically significant lower intercept than the reference group, as well as

Black and Hispanic students ($p < .001$). Students whose home language was not English also had a statistically significant lower intercept ($p < .001$). Socioeconomic status was positively significant as well, indicating that students from families with a higher SES were associated with higher science theta scores in the spring of Grade 2. Asian students and Indigenous/Multiracial students did not show statistically significant differences on intercept, as well as students with an IEP. The average rate of change for students was 0.05. Black, Hispanic, and Asian students showed statistically significant ($p < .001$) higher linear growth, as well as students whose home language was not English. Socioeconomic status was associated with statistically significant ($p < .001$) lower linear growth. The quadratic term results reveal that the reference group was estimated to have a statistically significant positive acceleration in science growth rate from Grade 1 to Grade 2 of 0.001. Black students showed statistically significant differences in rate of curvature ($p < .001$), with a decreased rate of -0.0004. The other statistically significant factor was socioeconomic status, with a positive acceleration of .0001 compared to the reference group. The addition of the student level predictors accounted for 26.43% of the variance in student intercepts, 32.70% of the variance in slopes, and 19.05% of the variance in curvature when compared to the unconditional model (see Table 9 for variance components, Pseudo-R^2, and deviance statistics). The deviance statistics calculated separately using each multiple imputation file show that the student level predictor model produced a statistically significant reduction in unexplained variation across each iteration ($p < .001$). Intercorrelations of the model parameters were -0.51 (intercept and linear), 0.62 (intercept and quadratic), and -0.84 (linear and quadratic).

The second conditional model included the addition of two student level predictors, initial reading and math theta scores from the fall of kindergarten baseline assessment (see column 3 in Table 8). The intercept for this model (M = 1.81) represents the mean science theta score in the spring of Grade 2 for White males whose home language is English and do not have an IEP across the Grade K – 2 time span. The same predictor variables were still significant as in the previous model, although the parameter estimates for the intercept decreased slightly for Black and Hispanic students, and students whose home language was not English. The variable for SES showed a more drastic decrease from the previous model (from 0.34 to 0.14). Both initial reading and math were statistically significant predictors (p < .001), indicating that, on average, students with higher initial math scores had science theta scores 0.34 points higher than the reference group and students with higher initial reading scores had science theta scores 0.16 points higher than the reference group. The average rate of change for students was 0.05, the same as found in the previous model. Asian, Black, and Hispanic students again showed statistically significant (p < .001) higher linear growth, as well as students whose home language was not English. Socioeconomic status was associated with statistically significant (p < .001) lower linear growth. The quadratic term results reveal that the reference group were estimated to have a statistically significant positive acceleration in science growth rate from Grade 1 to Grade 2 of 0.0001, a decrease from the first conditional model. Black students did not have statistically significant differences in rate of curvature in this model, but socioeconomic status remained significant, with a positive acceleration of .0001 compared to the reference group. With the addition of the prior achievement predictors of math and reading theta scores, the model covariates accounted for 48.53% of the variance in student intercepts, 48.88% of the variance in slopes, and 36.51% of the variance in curvature when compared to the unconditional model (see Table 9 for variance components, Pseudo-R^2, and deviance statistics). The deviance statistics show that the inclusion of the additional predictors in the model produced a statistically significant reduction in unexplained variation across each iteration ($p < .001$). Intercorrelations of the model parameters were -0.48 (intercept and linear), 0.60 (intercept and quadratic), and -0.83 (linear and quadratic).

Figure 2 shows the estimated science achievement growth trajectories for White males as compared to females, using the model-based parameter estimates when home language, IEP status, SES status, and initial reading and math status are controlled for. As shown, science achievement growth for females, while characterized by some acceleration in growth between kindergarten and Grade 2, overall exhibits a lower growth trajectory across the assessment time periods.

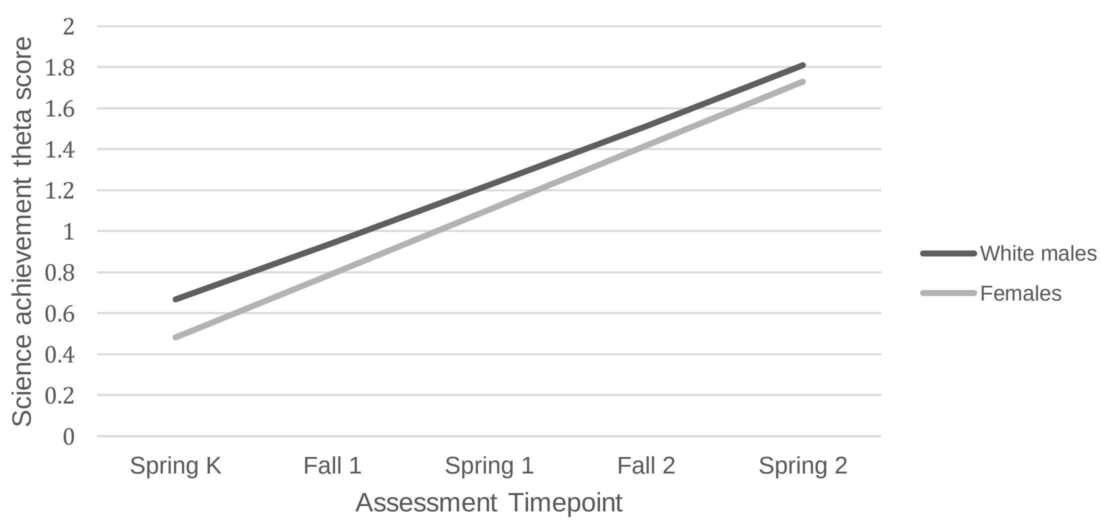


Figure 2. Estimated science achievement growth trajectories by gender

Figure 3 displays the estimated science achievement growth trajectories for White males as compared to students from other race/ethnicity categories examined in this study, using the model-based parameter estimates when home language, IEP status, SES status, and initial reading and math status are controlled for. Of particular note is the compensatory growth trajectory of Asian students, as well as the lower growth trajectory overall for Black students.

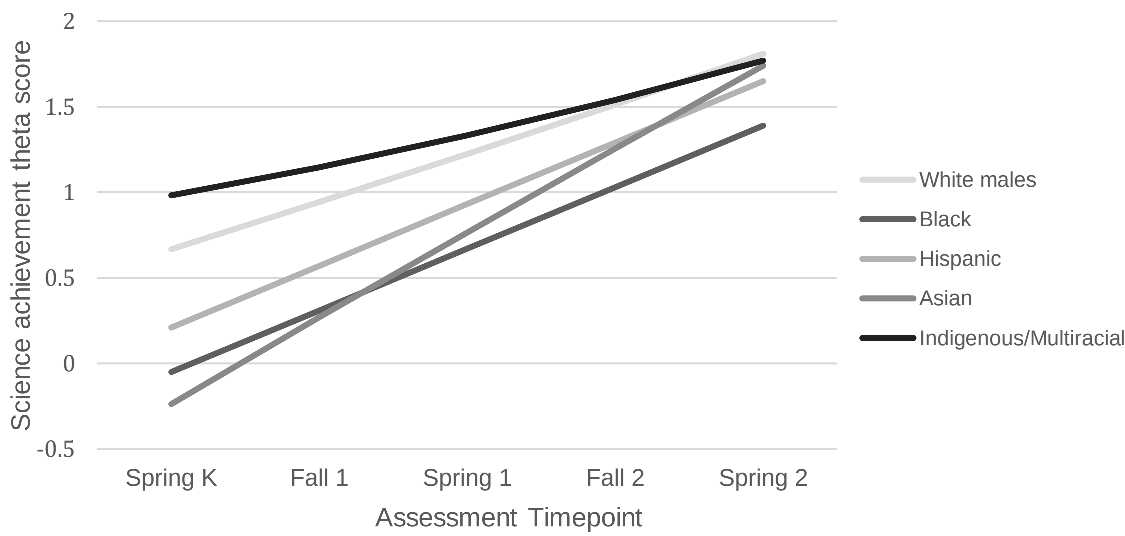


Figure 3. Estimated science achievement growth trajectories by race/ethnicity

The final conditional model included an average science instructional time predictor variable. The instructional variable was not statistically significant and caused almost no change to the parameter estimates for the intercept, slope, or curvature terms across groups (see column 4 in Table 8). Variance components and deviance statistics indicate almost no change from the previous model (see Table 9 for variance components, Pseudo-R^2, and deviance statistics).

Table 8

Fixed Effects for Longitudinal HLM Regression Models, Grades K - 2

	Unconditional			+Demographics			+Prior Reading and Math Achievement			+Instructional Time		
Fixed effect	Intercept	Linear	Quadratic	Intercept	Linear	Quadratic	Intercept	Linear	Quadratic	Intercept	Linear	Quadratic
Constant	**1.62**	**0.06**	**0.0004**	**1.86**	**0.05**	**0.001**	**1.81**	**0.05**	**0.0001**	**1.81**	**0.05**	**0.0005**
	(0.01)	(0.00)	(0.00)	(0.01)	(0.00)	(0.00)	(0.01)	(0.00)	(0.00)	(0.01)	(0.00)	(0.00)
Female				**-0.08**	0.0001	-0.0001	**-0.08**	0.002	-0.0001	**-0.08**	0.00	-0.0002
				(0.02)	(0.00)	(0.00)	(0.01)	(0.00)	(0.00)	(0.02)	(0.00)	(0.00)
Black				**-0.54**	**0.01**	**-0.0004**	**-0.42**	**0.01**	-0.0001	**-0.42**	**0.01**	-0.0003
				(0.03)	(0.00)	(0.00)	(0.03)	(0.00)	(0.00)	(0.03)	(0.00)	(0.00)
Hispanic				**-0.27**	**0.01**	-0.0001	**-0.16**	**0.01**	-0.0001	**-0.16**	**0.01**	-0.0001
				(0.03)	(0.00)	(0.00)	(0.02)	(0.00)	(0.00)	(0.02)	(0.00)	(0.00)
Asian				-0.01	**0.02**	-0.0002	-0.07	**0.03**	**-0.0002**	-0.08	**0.03**	**-0.00**
				(0.04)	(0.00)	(0.00)	(0.04)	(0.00)	(0.00)	(0.04)	(0.00)	(0.00)
Indigenous /Multiracial				-0.08	-0.01	0.0002	-0.04	-0.01	0.0002	-0.04	-0.005	0.0003
				(0.04)	(0.00)	(0.00)	(0.03)	(0.00)	(0.00)	(0.03)	(0.00)	(0.00)
Home language not English				**-0.23**	**0.01**	0.0001	**-0.15**	**0.01**	0.0001	**-0.15**	**0.01**	0.0001
				(0.03)	(0.00)	(0.00)	(0.02)	(0.00)	(0.00)	(0.03)	(0.00)	(0.00)
IEP Ever				0.02	0.002	-0.0001	0.04	0.003	-0.0001	0.04	0.001	-0.0001
				(0.02)	(0.00)	(0.00)	(0.02)	(0.00)	(0.00)	(0.02)	(0.00)	(0.00)
Composite SES measure				**0.34**	**-0.01**	**0.0001**	**0.14**	**-0.001**	**0.0001**	**0.14**	**-0.003**	**0.0001**
				(0.01)	(0.00)	(0.00)	(0.01)	(0.00)	(0.00)	(0.01)	(0.00)	(0.00)
Fall K reading score							**0.16**	**-0.01**	**0.0001**	**0.16**	**-0.01**	**0.0003**
							(0.01)	(0.00)	(0.00)	(0.02)	(0.00)	(0.00)
Fall K math score							**0.35**	-0.002	0.0001	**0.35**	-0.003	0.0001
							(0.02)	(0.00)	(0.00)	(0.02)	(0.00)	(0.00)
Science instructional time										0.0001	0.00002	-0.000001
										(0.00)	(0.00)	(0.00)

Note. Estimates are adjusted to account for the complex sampling design of the Early Childhood Longitudinal Study, Kindergarten Class of 2010–2011 (ELCS-K:2011) using sample weights and design effects. Standard errors are in parentheses. Bolding indicates statistical significance.

Table 9

Random Effects for Longitudinal HLM Regression Models, Grades K – 2

Random effect	Unconditional			+Demographics			+Prior Reading and Math Achievement			+Instructional Time		
	Intercept	Linear	Quadratic	Intercept	Linear	Quadratic	Intercept	Linear	Quadratic	Intercept	Linear	Quadratic
Variance component	**0.64**	**0.0006**	**0.00000**	**0.47**	**0.0005**	**0.00000**	**0.33**	**0.0005**	**0.00000**	**0.33**	**0.0005**	**0.00000**
Residual	0.14			0.14			0.14			0.14		
Pseudo R^2				26.43%	32.70%	19.05%	48.58%	48.88%	36.51%	48.59%	48.80%	36.51%
Deviance, File1			868.48 (4, <.001)			894.28 (12, <.001)			909.52 (14, <.001)			912.97 (15, <.001)
Deviance, File2			868.48 (4, <.001)			894.27 (12, <.001)			915.72 (14, <.001)			919.23 (15, <.001)
Deviance, File3			868.48 (4, <.001)			892.51 (12, <.001)			909.06 (14, <.001)			912.97 (15, <.001)
Deviance, File4			868.48 (4, <.001)			894.28 (12, <.001)			915.72 (14, <.001)			919.23 (15, <.001)
Deviance, File5			868.49 (4, <.001)			893.06 (12, <.001)			913.77 (14, <.001)			917.47 (15, <.001)

Note. Estimates are adjusted to account for the complex sampling design of the Early Childhood Longitudinal Study, Kindergarten Class of 2010–2011 (ELCS-K:2011) using sample weights and design effects. Bolding indicates statistical significance.

A final model was run including the specific interaction terms to answer research question six, even though the results indicated no statistical significance for the science instructional time variable. This model is presented separately in Table 10. These results are taken up more thoroughly in the Discussion section.

Table 10

Fixed and Random Effects for Longitudinal HLM Regression Model Including Instructional Time and Interaction Terms, Grades K – 2

	+Instructional Time and Interactions		
Fixed effect	Intercept	Linear	Quadratic
Constant	**1.82** (0.01)	**0.05** (0.00)	**0.0006** (0.00)
Female	**-0.11** (0.02)	0.002 (0.00)	-0.0002 (0.00)
Black	**-0.48** (0.03)	0.008 (0.00)	-0.0003 (0.00)
Hispanic	**-0.19** (0.03)	**0.01** (0.00)	-0.0001 (0.00)
Asian	-0.07 (0.06)	**0.03** (0.00)	**-0.00** (0.00)
Indigenous /Multiracial	-0.07 (0.05)	-0.003 (0.00)	0.0001 (0.00)
Home language not English	**-0.16** (0.03)	**0.01** (0.00)	0.0001 (0.00)
IEP Ever	0.03 (0.02)	0.001 (0.00)	0.0001 (0.00)
Composite SES measure	**0.14** (0.01)	**-0.003** (0.00)	**0.0001** (0.00)
Fall K reading score	**0.16** (0.01)	**-0.01** (0.00)	**0.0003** (0.00)
Fall K math score	**0.35** (0.02)	-0.003 (0.00)	0.0001 (0.00)
Science instructional time	0.0001 (0.00)	0.00002 (0.00)	-0.000001 (0.00)
Female X Black	0.07 (0.05)	0.002 (0.01)	0.0001 (0.00)
Female X Hispanic	0.06 (0.04)	-0.004 (0.00)	0.0002 (0.00)
Female X Asian	0.001 (0.07)	-0.01 (0.01)	0.0002 (0.00)
Female X Indigenous/Multiracial	0.06 (0.07)	-0.005 (0.01)	0.0002 (0.00)

	+Instructional Time and Interactions		
Random effect	Intercept	Linear	Quadratic
Variance component	**0.33**	**0.0005**	**0.00000**
Residual	0.14		
Pseudo R^2	47.59%	47.80%	35.51%
Deviance, File1			1071.91 (19, <.001)
Deviance, File2			1071.36 (19, <.001)
Deviance, File3			1070.23 (19, <.001)
Deviance, File4			1071.91 (19, <.001)
Deviance, File5			1071.90 (19, <.001)

Note. Estimates are adjusted to account for the complex sampling design of the Early Childhood Longitudinal Study, Kindergarten Class of 2010–2011 (ELCS-K:2011) using sample weights and design effects. Bolding indicates statistical significance.

Effect Size

Using the parameters of the growth model, science achievement growth effect sizes were calculated to provide an empirical context for interpretation of differences in student growth across years. Table 10 shows the growth effect sizes by student gender and race/ethnicity categories, where the estimated growth rate at each assessment time point serves as the metric of the magnitude of change (Stevens, 2018). As shown in the table, student growth rates decreased slightly from the first assessment occasion in the spring of kindergarten to the final assessment the spring of Grade 2. Cohen's (1988) rules of thumb for interpretation of effect sizes (i.e., 0.2, 0.5, and 0.8 interpreted as small, medium, and large effects) were not designed for growth effect size measures. Rather, the relative differences both across years for all students and in comparisons between groups of students should be used in examining the magnitude of growth effect size (Bloom et al., 2008). The growth effect size from kindergarten to Grade 1 is characterized by large effect sizes for all student groups, ranging from 0.87 for Black students to 0.95 for Asian students. Males and females had similar growth effect sizes, at 0.90 and 0.88

respectively. The annual science achievement growth effect size from Grade 1 to Grade 2 was lower, although still moderate to large in terms of effects, for most student groups, ranging from 0.71 for Indigenous/Multiracial students to 0.88 for Asian students. However, the growth effect size for Black students increased slightly, from 0.87 to 0.88. The difference in effect size between males and females widened in Grade 2, at 0.82 and 0.78 respectively.

Table 11

Science Achievement Growth Rate at Each Assessment Time Point by Student Group and Annual Growth Rate Effect Sizes

Student Group	Kindergarten Spring	Grade 1 Fall	Grade 1 Spring	Grade 2 Fall	Grade 2 Spring	Annual Growth ES K – 1	Annual Growth ES 1 – 2
All students	0.067	0.072	0.067	0.062	0.060	0.91	0.83
Gender							
Female	0.066	0.071	0.065	0.060	0.059	0.88	0.78
Male	0.065	0.074	0.066	0.059	0.067	0.90	0.82
Race/ethnicity							
Asian	0.065	0.079	0.071	0.064	0.062	0.95	0.82
Black	0.071	0.063	0.065	0.067	0.068	0.87	0.88
Hispanic	0.068	0.070	0.067	0.066	0.066	0.90	0.86
Indigenous /Multiracial	0.065	0.074	0.065	0.055	0.054	0.90	0.71
White	0.065	0.075	0.065	0.060	0.051	0.90	0.72

Note. Empirical Bayes estimates were used in calculating growth effect size.

CHAPTER V

DISCUSSION

In this section, the results of the analyses conducted in this study are examined in relation to the specific research questions posed at the outset. Limitations to the study are discussed, as well as possibilities for future avenues of explorations.

Growth Trajectories

To address the first two research questions, the results of this study show that for the ECLS-K: 2011 science achievement measure, achievement growth over Grades K to 2 was best represented as a curvilinear function with achievement growth slightly accelerating over time. While studies in other content areas such as reading and math generally suggest deceleration in achievement growth over time (e.g., Morgan, Farkas, & Wu, 2009), the limited grade span of the present study is likely why, on average, growth is still accelerating at Grade 2. However, Morgan and colleagues (2016), looked at science achievement growth for students in Grades 3 – 8 and found that the growth trajectories for students across those grades, while not exhibiting a strong curvilinear positive acceleration, continued to grow without exhibiting the characteristic deceleration often noted in studies of learning trajectories. The initial level of science achievement across all students appears to be highly predictive of growth and science achievement, corresponding to an initial hypothesis that one would expect to see students with lower initial levels of science achievement experience lower growth. However, variation in science achievement growth is explained in part by several other factors that are further explored in the subsequent research questions. Additionally, the widening range of science achievement scores in Grade 2 indicate that while overall growth trajectories may remain stable over time, an increase in the variability of science scores between students may mean that some students begin to fall behind their peers in terms of average achievement, while others make greater strides. It is important to note that an individual outlier low score may not correspond directly to a student's actual ability (e.g., bad testing day, interrupted testing session, etc.). However, a further examination of the subgroups of students whose science scores fall outside of an established average range by grade level may provide interesting insights into overall science growth patterns.

Even without the inclusion of additional explanatory factors, there are implications for science education policy here. The robust growth demonstrated in science achievement in the early elementary years suggests that these years may be critical to the development of foundational attitudes and interest in science. A developmental phenomenon known as the "five to seven year shift" posits that children experience a transition around this age, when most children are in the pre-school to first grade range, characterized by increased abilities for self-regulation, complex thought processing, and cooperative interpersonal relationships (Weisner, 1996). Recent studies have shown that even very young children engage in causal reasoning and statistical learning, two skills that are associated with the development of scientific inquiry thinking skills (Hadani & Rood, 2018). Additionally, in studies investigating the onset of initial interest in STEM, individuals that end up following STEM-related career pathways frequently report that the early childhood and elementary years were critical to stimulating the interest that led to continued engagement with STEM activities (Maltese, Melki, & Wiebke, 2014).

To capitalize on these factors, policymakers and early educators should be encouraged to systematically and explicitly engage all young children in activities designed to promote STEM-related thinking skills, even before they start school.

Demographic Predictor Variables

The addition of student level predictors in the first conditional growth model allowed for the examination of children's gender, race/ethnicity, family SES status, home language, and IEP status with respect to kindergarten to second grade growth trajectories in science, the goal of the third research question. As with other recent studies of science achievement in the early elementary years (Curran & Kellogg, 2016), statistically significant differences in science achievement by both gender and race/ethnicity were found to develop earlier than previous research had suggested, although this gap in the literature may be due to the fact that most large scale assessments of science achievement do not happen until the 3rd or 4th grade in most states. Using the same dataset as was used in the present study, Curran and Kellogg's work showed no gender gap in Kindergarten on the science measure. However, a statistically significant gender gap was detected in Grade 1, indicating that, for gender at least, students may enter Kindergarten on fairly equal standing. The results of this study showed a statistically significant difference in science achievement by gender in Grade 2, linking Curran and Kellogg's results and the results of other researchers examining science achievement gender gaps from Grades 3 – 8 (Quinn & Cooc, 2015; Morgan et al., 2016). Results from these studies suggest that the gender gap persists across grade levels, and, after the initial widening, maintains a stable trajectory, as was seen between Grades K – 2 in the present study. Given that the gender gap in science does not appear until after the onset of students' K – 12 schooling and that the results of many studies suggest that female students start to lose interest in STEM classes or career-related pathways in middle school (NSB, 2016), it does not make sense to wait until middle or high school to intervene with strategies designed to help encourage girls in science. One study that looked at girls' interest in STEM found that females were more likely to report that their initial interest in STEM arose through a class at school than males (Maltese & Cooper, 2017). Additionally, this same study found that females were more likely to report that their teachers had an influence on their STEM interest, suggesting the importance of the elementary school teacher's role in engaging girls with science at a young age. Previous research has shown that gender differences in attitude around perceived self-confidence in science begins in the early elementary years (Andre, Whigham, Hendrickson, & Chambers, 1999), and that parents echoed similar attitudes about the importance of science for boys over girls. Given the role that self-efficacy and attitude about science has been shown to play in girls' interest and engagement with science, parents should also be considered when planning interventions and supports. A fascinating new meta-analysis examined the results of over 5 decades of Draw-A-Scientist Test studies and found that while children in more recent studies more frequently produced drawings of women as scientists than in older studies using this test, on average, even in the more recent studies, scientists were more likely to be depicted as men (Miller, Nolla, Eagly, & Uttal, 2018). This suggests that while the representation of women has increased in the sciences in general, this visibility overall still may not be as apparent to children, potentially having downstream effects on girls' interest and engagement with science.

When examining science achievement growth in this study by race/ethnicity, the results show that at the time of the first kindergarten science assessment, there were already differences between groups. Curran and Kellogg found significant gaps in science achievement between White students and both Black and Hispanic students. At Grade 2, these gaps remained and, similar to the growth pattern seen with the other studies of science growth (Quinn & Cooc, 2015; Morgan et al., 2016), appear to be fairly stable across time for Hispanic students, but slightly increasing for Black students, suggesting the presence of the Matthew effect in science achievement growth. Also of interest is the way in which the Asian students' growth, in relation to White students, appears to have a compensatory trajectory, such that there is not a statistically significant difference between White and Asian student science scores at the end of Grade 2 assessment. Indigenous/Multiracial students (including Native Hawaiian/Other Pacific Islander, American Indian or Alaska Native, More Than One Race) did not have a statistically significant difference from the reference group on the science achievement measure at Grade 2. Given that this category includes students from several different subgroups, future research that allows for the specific examination of each student group is warranted, to determine if the non-significant difference was a biased estimate based on grouping.

In general, and in keeping with the research on the effects of family socioeconomic status on student achievement across grade levels and content areas, SES was found to be a significant predictor of science achievement at Grade 2. While the inclusion of the SES variable helps to explain some of the variability in science achievement at Grade 2 for Black and Hispanic students, it did not explain it entirely.

Additional factors beyond socioeconomic status can therefore be attributed to the differences in science achievement by student group. As noted previously in the literature review, student level factors associated with SES, such as lack of family resources allowing students to access science related resources at home or in informal/after school settings and the availability of parental time or awareness for encouragement of science participation, have been cited as possible contributors to differences in achievement. However, the effects of inequities in neighborhood and school level socioeconomics and resource allocation on student achievement in science, especially for Black and Hispanic students, has been well-documented in other research (e.g., Museus et al., 2011). Quinn (2015), in a recent study on the effects of SES on achievement gaps in reading and math, found that school quality helped to explain widening achievement gaps in reading for black students over kindergarten. Additionally, Quinn and Cooc (2015) found that classroom level variables were important contributors to science achievement gaps for Black and Hispanic students across the span of Grades 3 to 8. While this study did not focus on classroom level or school level factors beyond time spent on science instruction, the amount of unexplained variability and the results of other research suggests that there are many questions left unanswered about the specific mechanisms involved to understand the effects of SES on student achievement and growth, as this study was limited by its observational scope.

Student disability status, as indicated by whether or not students had an Individualized Education Plan at any time across the three year span of the data used in this study, was included in the set of student demographic variables, along with an indicator for the home language of the participating student. These variables were included as they were hypothesized to potentially have an effect on science achievement. Results of the analyses indicate that there were no significant differences in science achievement at Grade 2 for students with an IEP as compared to the reference group. Given the small grade range of the data in this study and the fact that some students may not yet have been identified with a learning disability covered by an IEP, this finding may not be surprising. However, future studies of any group differences in early science achievement for students with disabilities could be further investigated by drilling down into the data and parsing out specific exceptionality categories for analysis.

Student home language was, however, a significant predictor of science achievement. Students that were reported as having a home language that was not English had statistically significant lower science scores than the reference group at the Grade 2 assessment. Depending on the model of instruction and the degree of services students who are developing proficiency with English are receiving at school, it is possible that these students are faced with challenges in learning from early science instruction. It is important to note that the ECLS-K assessments accommodated students who were Spanish-language speakers in kindergarten and Grade 1, but that all assessments in Grade 2 were administered in English. Students who did not speak English or Spanish at home did not have alternate test versions; if they did not pass the English language screener and they did not speak Spanish, they were excluded from the assessments and only provided their height and weight during the assessment period. Additionally, students who received the Spanish language version of the assessments in kindergarten and Grade 1 did not complete a science assessment, as there was no Spanish language version provided. This exclusion rule may have produced biased estimates for students who passed the

English language screener but were still developing proficiency with the English language. While Curran and Kellogg (2016) did not include a home language variable in their analysis using this same dataset, Morgan and colleagues (2016) identified lower science achievement scores for students in Grade 3 from non-English speaking homes, using the previous version of the ECLS-K data, that continued through Grade 8, although the gap decreased over time. They hypothesize that the narrowing of the achievement gap between students whose home language is English and those from non-English speaking homes may in part be explained by the fact that as students age, they gain more proficiency with English. However, for students who do not receive adequate services to support their language development, there is the possibility that their science achievement growth will not exhibit this compensatory effect, suggesting the need for further investigation into the specific mechanics at play for students whose home language is not English.

Prior Reading and Math Achievement

To examine the fourth research question, student reading and math achievement in the fall of kindergarten were added to the model to determine how prior achievement in reading and math helped to explain kindergarten to second grade science growth and achievement at Grade 2. Both continuous variables were statistically significant predictors of science achievement, although they acted in slightly different ways. Fall of kindergarten math scores were associated with a higher magnitude of change on the intercept (b = 0.35) than fall of kindergarten reading scores (b = 0.16); however, the reading scores were also associated with a statistically significant decrease in linear change (b = -0.01) and increase in quadratic curvature (b = 0.0001) at Grade 2. The inclusion of initial math scores as a predictor did not change either linear or quadratic terms. The addition of initial reading and math scores did not change the intercept coefficient for females, which remained significantly lower than the reference group. However, the intercept coefficient for Black and Hispanic students decreased, as well as the intercept coefficients for student home language family status and for SES, indicating that prior reading and math skill helps to parse out and explain some, but not all, of the variability associated with most of the significant model covariates. The research on the effects of prior reading and math skills on science achievement is fairly well established for older students. Quinn and Cooc (2015) reported that prior reading and math achievement had differential effects in explaining science achievement in their study with students in Grades 3 – 8. Their findings suggested that prior math achievement did more to explain the science achievement gap between White and Black students than did reading, while, conversely, prior reading achievement did more to explain the science achievement gap between White and Hispanic students. The results of this study indicate that a similar relation between early reading and math achievement is related to science achievement in the early elementary school years. Given that these differences begin early and have been shown to persist into middle and high school, it is not enough to assume that as elementary school students develop their reading and math skills over time that corresponding gaps in science achievement will decrease. The increasing complexity of science vocabulary and content area reading and math skills needed as students progress to later grades (Flores, 2007; Snow, 2010), often coupled with a decrease in the amount of direct support for students who are less than proficient with these skills, suggests that further inquiry is needed into the interplay between reading, math, and science achievement. Additionally, the effects of socioeconomic status, both at the student and school levels, are deeply interwoven into the patterns of achievement across the content areas and how they interact with one another.

Science Instructional Time

The final research questions guiding this study examined the inclusion of a classroom level variable, the average amount of science instructional time students in the study sample experienced across the K – 2 grade span. Given the high degree of variability in science instructional time reported by participating study teachers and previous research linking decreased science instructional time with a decline in science achievement scores on the NAEP 2009 4th grade science assessment (Blank, 2013), it seemed reasonable to speculate that students receiving less direct science instruction in the early grades would have lower science achievement scores. However, the results of the analysis in this study indicated no statistically significant effect of average science instruction time on science achievement or growth for any group of students, nor did the inclusion of the specific interaction terms identified in the last research question (i.e., the examination of the moderating effect of an interaction between female and race/ethnicity status on science instructional time) result in statistical significance.

The results of this study are replicated in very recent work by Curran and Kitchin (2018), also using the ECLS-K: 2011 dataset, that examined whether the amount of time spent on science instruction, calculated at minutes per week, predicted science achievement in the spring of kindergarten, Grade 1, and Grade 2. They used a series of OLS regressions to predict science achievement, with student, family, teacher, and school background control variables. Their analyses showed that time did not consistently predict higher science achievement scores.

The literature base regarding the effects of instructional time is mixed and complex. There are examples of studies that find positive associations and increases in academic achievement with increased, content-specific instructional time (e.g., Andersen, Humlu, & Nandrup, 2016) while other studies find no significant increases in achievement due to increased instructional time (e.g., Baker, Fabrega, Galindo, & Mishook, 2004). Given the inconsistent findings and possibility of content- or population-specific differences in effect, it may be more important to focus more on the method and quality of instructional time in elementary science. The descriptive nature of the present study is limited to the available data; to better understand the relation between instructional time and science achievement, experimental methodologies are likely better suited to detecting salient differences. Although instructional time as it was defined in this study may not have been significant, there are other aspects of instructional time that may require a more nuanced and comprehensive approach.

Science Achievement Growth Effect Size

Bloom and colleagues (2008) demonstrated the importance of deriving interpretations of the effect of an intervention (in this case, the normal growth experienced by a student in a content area from year to year, which includes any school based interventions and other developmental factors that guide student achievement growth) in the context of expectations for the grade or grades of interest. In this study, the use of a longitudinal design dictated the use of a growth effect size to take into account the parameters of the estimated growth model (Stevens, 2018). The results in general show large year-to-year effect sizes in both the Grades K – 1 and Grades 1 – 2 time spans, with differences appearing by student group. By race/ethnicity, for Grades K – 1, Asian students exhibited the largest effect size, while Black students exhibited the lowest, although they were both still relatively large in terms of effect size interpretation. While Asian, Hispanic, White, and Indigenous/Multiracial students experienced a decrease in growth effect size in the Grade 1 – 2 time span, Black students exhibited a slight increase in growth effect size in this time span. By gender, females exhibited a slightly smaller growth effect size as compared to males in Grades K – 1 that increased in the magnitude of the difference in the Grade 1 – 2 time span.

The results of the effect size calculations appear to correspond to the strong pattern of science achievement growth detected in this study's analyses across the first three years of elementary school for this sample of students. It seems reasonable to expect that students who are being introduced to science instruction, many possibly for the first time, would experience this type of growth as a result of typical instruction, although the author was unable to locate any comparison growth effect size results for science or any other content areas in the same grade range. It is possible that differences in effectiveness of science instruction for different subgroups could be detected, using groupings based on other student level characteristics or other teacher/classroom/school level variables. The growth effect sizes presented in this study may be a unique contribution for future examinations and comparisons of science achievement growth in the early elementary years. Of particular interest would be to continue to follow this cohort of students with the release of the remaining data collected in the ECLS-K:2011 study to examine the year-to-year growth patterns by subgroup through Grade 5.

Limitations

The present study is descriptive and intended to describe and explore potential relationships mainly between science achievement and student demographic characteristics. While the findings from the current study offer some exciting new directions and re-directed focus in the understanding of science achievement across the K-12 student experience, several limitations must be noted. For one, the current study made use of a large, nationally representative dataset, which offered statistical power and generalizability, as well as the rare opportunity to directly examine longitudinal science achievement in the elementary grades. However, the use of an extant, observational dataset limits the measures and indicators that are available for the constructs and variables of interest. There may be more science assessment tasks that would more comprehensively represent science achievement, and the ECLS-K is entirely missing student self-efficacy survey items targeted directly at experience in science below Grade 3, which is an important limitation to the present study. In addition to this limitation, there are a number of additional relevant variables in the ECLS-K dataset that, if used in the analyses, may have provided more or better information, including other family characteristics, specific teacher and classroom levels instructional variables, and school type and location. Additionally, the lack of access to the actual assessment measures and a more complete description of the psychometric properties of the measures is potentially problematic.

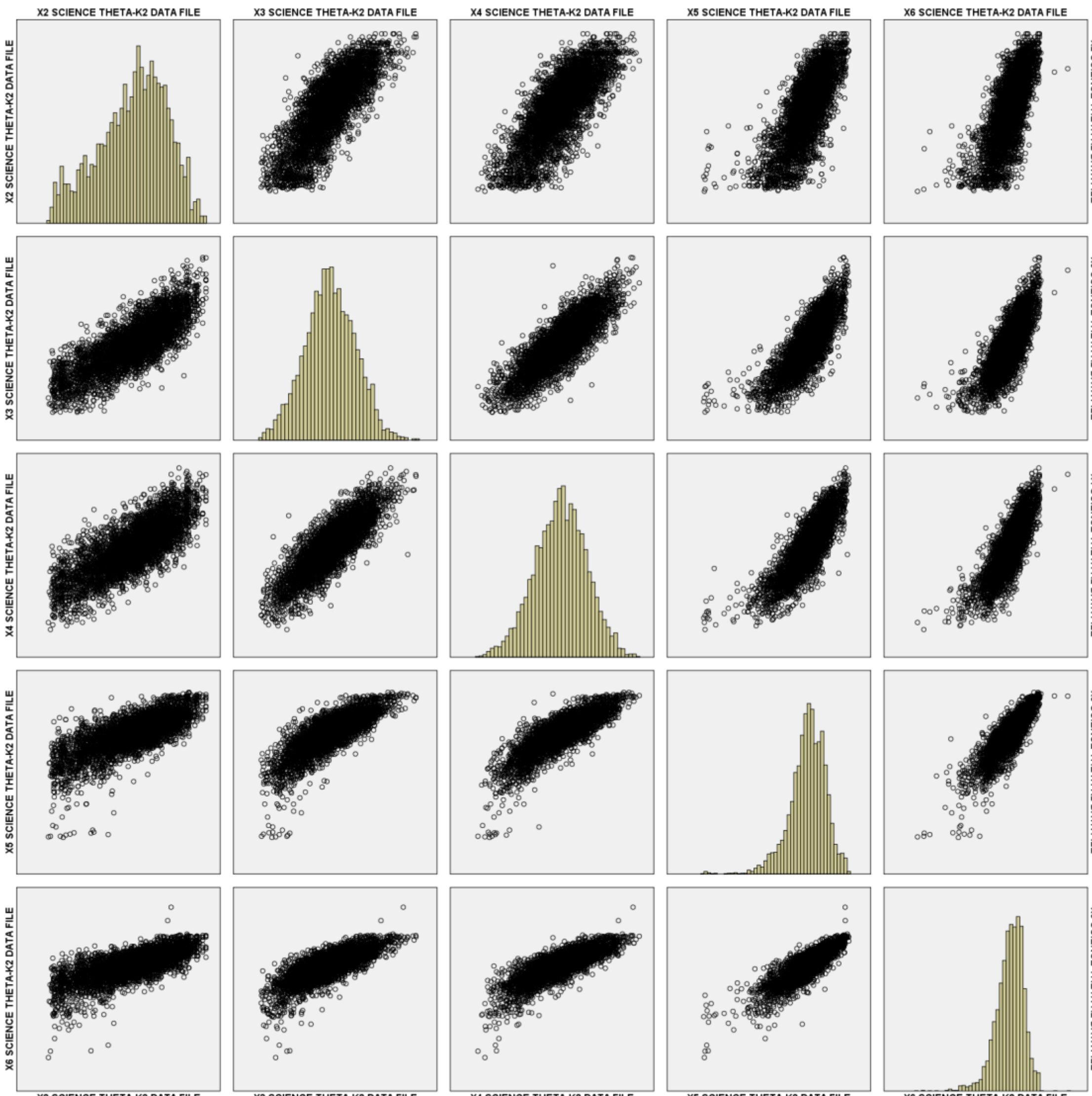
Despite the redesign of the ECLS-K: 2011 study, there are also data collection issues that limit the generalizations that can be drawn from these analyses. While sampling weights were designed to help account for the degree of attrition and missingness seen in this dataset, there is still the potential for bias introduced by the attrition. The missingness analyses conducted in the present study indicate patterns of missing data. While full information maximum likelihood combined with multiple imputation is a strong approach to dealing with the missing data, additional analyses of the missingness patterns may reveal methodological approaches that would result in more precise estimation. Additionally, the wide windows of administration of the cognitive assessments in the fall and spring of each year, producing inconsistent periods of time between assessments for individual students, may have led to biased estimates of student science achievement and growth with the methodology that was used in this study. More appropriate techniques for handling the assessment administration discrepancies for students may produce better estimates.

Conclusion

The findings of this study contribute to a very small, but growing and important, body of research on science achievement and growth in the early elementary years. While efforts to motivate and engage middle and high school age students in science should remain in place and continue to expand, understanding the beginnings of disparities in science achievement may help to put into place practices that can be used to scaffold and build early science learning opportunities. Knowing that gaps in science achievement and growth start early and persist over time lends an urgency to capitalizing on young students' innate interest in the world around them. Recent developments in our understanding of cognitive development in pre-school age children indicate that by early elementary school, children can be ready to begin to use scientific thinking practices that are often not introduced until later in their educational experience. Making science real and relevant and visible, introducing students to a variety of role models and pathways that allow them to see themselves in the world of science, and increasing students' science self-efficacy beliefs at a young age are just a few examples of promising practices that need to be better explored in the early elementary years. While increasing and diversifying the number of students that pursue and persist in STEM-related college and career pathways is a laudable goal for both economic and innovation-building purposes, we also need to continue to grow a scientifically literate society. The kinds of critical thinking, questioning, and communication skills that participation in higher level STEM classes help to promote are essential for all citizens of an increasingly connected and technocentric world.

APPENDIX

HISTOGRAMS AND SCATTERPLOTS OF SCIENCE ACHIEVEMENT SCORES ACROSS ASSESSMENT TIME POINTS



REFERENCES CITED

American Association for the Advancement of Science. (1994). *Benchmarks for science literacy*. Oxford University Press.

Andersen, S. C., Humlum, M. K., & Nandrup, A. B. (2016). Increasing instruction time in school does increase learning. *Proceedings of the National Academy of Sciences of the United States of America*, *113*(27), 7481–7484. http://doi.org/10.1073/pnas.1516686113

Andre, T., Whigham, M., Hendrickson, A., & Chambers, S. (1999). Competency beliefs, positive affect, and gender stereotypes of elementary students and their parents about science versus other school subjects. *Journal of Research in Science Teaching*, *36*(6), 719-747.

Appleton, K. (2006). Science pedagogical content knowledge and elementary science teachers. In K. Appleton (Ed.), *Elementary science teacher education: International perspectives on contemporary issues and practice* (pp. 31 – 54). Mahweh, NJ: Lawrence Erlbaum Associates.

Archer, L., DeWitt, J., Osborne, J., Dillon, J., Willis, B., & Wong, B. (2012). Science aspirations, capital, and family habitus: How families shape children's engagement and identification with science. *American Educational Research Journal*, *49*(5). 881 – 908.

Atkinson, R.D., & Mayo, M. (2010). *Refueling the U.S. innovation economy: Fresh approaches to science, technology, engineering and mathematics (STEM) education*. Washington, DC: Information Technology and Innovation Foundation. http://www.itif.org/publications/refueling-us-innovation-economy-fresh-approaches-stem-education. Accessed August 2016.

Augustine, N. R. (2005). *Rising above the gathering storm: Energizing and employing America for a brighter economic future*. National Academies Press.

Augustine, N. R., Barrett, C., Cassell, G., Grasmick, N., Holliday, C., & Jackson, S. A. (2010). *Rising above the gathering storm, revisited: Rapidly approaching category 5.* Washington, DC: National Academy of Sciences, National Academy of Engineering, Institute of Medicine.

Baker, D. P., Fabrega, R., Galindo, C., & Mishook, J. (2004). Instructional time and national achievement: Cross-national evidence. *Prospects, 34*(3), 311 – 334.

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*. Chapel Hill, NC: Horizon Research, Inc.

Berliner, D. C. (1990). What's all the fuss about instructional time? *The nature of time in schools: Theoretical concepts, practitioner perceptions*, 3-35.

Berryman, S. E. (1983). *Who will do science? Trends, and their cause in minority and female representation among holders of advanced degrees in science and mathematics*. New York, NY: Rockefeller Foundation.

Bian, L., Leslie, S. J., & Cimpian, A. (2017). Gender stereotypes about intellectual ability emerge early and influence children's interests. *Science*, *355*(6323), 389-391.

Biesanz, J. C., Deeb-Sossa, N., Papadakis, A. A., Bollen, K. A., & Curran, P. J. (2004). The role of coding time in estimating and interpreting growth curve models. *Psychological Methods*, *9*(1), 30 - 52.

Blank, R. K. (2013), Science instructional time is declining in elementary schools: What are the implications for student achievement and closing the gap? *Science Educator*, *97*, 830–847

Bloom, H. S., Hill, C. J., Black, A. R., & Lipsey, M. W. (2008). Performance trajectories and performance gaps as achievement effect-size benchmarks for educational interventions. *Journal of Research on Educational Effectiveness*, *1*(4), 289-328.

Bracey, G.W. (2000). The TIMSS 'final year' study and report: A critique. *Educational Researcher 29*(4), 4–10.

Breiner, J.M., Harkness, S.S., Johnson, C.C., & Koehler, C.M. (2012). What is STEM? A discussion about conceptions of STEM in education and partnerships. *School Sci. Math. 112*(1), 3–1.

Byrnes, J.P., & Miller, D.C. (2006). The relative importance of predictors of math and science achievement: An opportunity propensity analysis. *Contemporary Educational Psychology*, *32*, 599 – 629.

Cavanaugh, S. (2007). Federal rule yields hope for science. *Education Week*, *27*(7), 1–13.

Chiu, M. M., & Khoo, L. (2005). Effects of resources, inequality, and privilege bias on achievement: Country, school, and student level analyses. *American Educational Research Journal*, *42*(4), 575-603.

Clowse, B. B. (Ed.). (1981). *Brainpower for the Cold War: The Sputnik Crisis and National Defense Education Act of 1958* (No. 3). Greenwood Publishing Group.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Conley, D.T. (2007). *Redefining college readiness,* (Vol. 3). Eugene, OR: Educational Policy Improvement Center.

Cotton, K., & K. R. Wikelund. (2001). *Parent involvement in education. School Improvement Research Series.* Portland, OR: Northwest Regional Educational Laboratory. Retrieved from www.nwrel.org/scpd/sirs/3/cu6.html.

Culp, K. M., Honey, M., & Mandinach, E. (2005). A retrospective on twenty years of education technology policy. *Journal of Educational Computing Research, 32*(3), 279-307.

Curran, F. C., & Kellogg, A. T. (2016). Understanding science achievement gaps by race/ethnicity and gender in kindergarten and first grade. *Educational Researcher, 45*(5), 273-282.

Curran, F. C., & Kitchin, J. (2018, April). *Does more time on science or science/topics skills predict science achievement in the early grades?* Poster presented at the 2018 annual meeting of the American Educational Research Association, New York, NY.

Darolia, R., Koedel, C., Main, J. B., Ndashimye, F., & Yan, J. (2018). *High school course access and postsecondary STEM enrollment and attainment (Working Paper No. 186).* Retrieved from National Center for Analysis of Longitudinal Data in Education Research website: www.caldercenter.org

Duncan, R. G., Rogat, A., & Yarden, A. (2009). A learning progression for deepening students' understandings of genetics across the 5th - 10th grades. *Journal of Research in Science Teaching, 46*(6), 655 - 674.

Else-Quest, N. M., Mineo, C. C., & Higgins, A. (2013). Math and science attitudes and achievement at the intersection of gender and ethnicity. *Psychology of Women Quarterly, 37*(3), 293 – 309.

Enders, C. (2010). *Applied missing data analysis*. New York: Guilford Press.

Every Student Succeeds Act, S.1177, 114th Cong. (2015). Retrieved from http://edworkforce.house.gov/uploadedfiles/every_student_succeeds_act_-_conference_report.pdf

Flores, A. (2007). Examining disparities in mathematics education: Achievement gap or opportunity gap? *High School Journal, 91,* 29–42.

Good, C., Aronson, J., & Harder, J. A. (2007). Problems in the pipeline: Stereotype threat and women's achievement in high-level math courses. *Journal of Applied Developmental Psychology*, 29, 17–28.

Graham, J. W., Olchowski, A. E., & Gilreath, T. D. (2007). How many imputations are really needed? Some practical clarifications of multiple imputation theory. *Prevention Science*, *8*, 206-213.

Hahs-Vaughn, D. L. (2005). A primer for using and understanding weights with national datasets. *Journal of Experimental Education*, *73*(3), 221 - 248.

Harackiewicz, J. M., Rozek, C. S., Hulleman, C. S., & Hyde, J. S. (2012). Helping parents to motivate adolescents in mathematics and science: An experimental test of a utility-value intervention. *Psychological Science*, *23*(8), 899 – 906.

Herrera, F. A., Hurtado, S., Garcia, G. A., & Gasiewski, J. (2012). *A model for redefining STEM identity for talented STEM graduate students*. In Proc. of American Educational Research Association Annual Conference (pp. 13-17).

Honey, M., Pearson, G., & Schweingruber, H., eds. (2014). *STEM Integration in K-12 Education: Status, Prospects, and an Agenda for Research*. Washington, DC: Natl. Acad. Press.

IBM Corp. (2017). IBM SPSS statistics for Windows, Version 25.0. Armonk, NY: Author.

Julien, H. & Barker, S. (2009). How high-school students find and evaluate scientific information: A basis for information literacy skills development. *Library and Information Science Research,* 31, 12-17.

Kastberg, D., Roey, S., Ferraro, D., Lemanski, N., & Erberber, E. (2013). *U.S. TIMSS and PIRLS 2011 Technical Report and Users Guide.* (NCES 2013-046). U.S. Department of Education. Washington, DC: National Center for Education Statistics. Available from http://nces.ed.gov/pubsearch.

Kesidou, S., & Roseman, J.E. (2002). How well do middle school science programs measure up? Findings from Project 2061's curriculum review. *Journal of Research in Science Teaching, 39*, 522 – 549.

Kohlhaas, K., Lin, H. H., & Chu, K. L. (2010a). Disaggregated outcomes of gender, ethnicity, and poverty on fifth grade science performance. *RMLE Online, 33*(7), 1-12.

Kohlhaas, K., Lin, H. H., & Chu, K. L. (2010b). Science equity in third grade. *The Elementary School Journal, 110*(3), 393-408.

Koretz, D. (2008). *Measuring up: What educational testing really tells us*. Cambridge, MA: Harvard University Press.

Ladson-Billings, G. (2006). From the achievement gap to the education debt: Understanding achievement in U.S. schools. *Educational Researcher, 35*, 3–12.

Langdon, D., McKittrick, G., Beede, D., Khan, B., & Doms, M. (2011). *STEM: Good jobs now and for the future.* ESA Issue Brief# 03-11. US Department of Commerce.

Lin, H. H. (2014). The comparisons between ethnicity and poverty level on eighth graders' science performance. *Journal of Studies in Education*, *4*(4), 122-130.

Loughran, J., Mulhall, P., & Berry. A. (2004). In search of pedagogical content knowledge in science: Developing ways of articulating and documenting professional practice. *Journal of Research in Science Teaching, 41*(4), 370 – 391.

Maltese, A. V., & Cooper, C. S. (2017). STEM Pathways: Do men and women differ in why they enter and exit? *AERA Open*, *3*(3), DOI: 10.1177/2332858417727276.

Maltese, A. V., Melki, C. S., & Wiebke, H. (2014). The nature of experiences responsible for the generation and maintenance of interest in STEM. *Science Education, 98*(6), 937 – 962.

McClure, E. R., Guernsey, L., Clements, D. H., Bales, S. N., Nichols, J., Kendall-Taylor, N., & Levine, M. H. (2017). *STEM starts early: Grounding science, technology, engineering, and math education in early childhood.* New York: The Joan Ganz Cooney Center at Sesame Workshop.

Miller, D.I., Nolla, K.M., Eagly, A.H., & Uttal, D.H. (2018). The development of children's gender-science stereotypes: A meta-analysis of 5 decades of U.S. Draw-A-Scientist studies. *Child Development, 00*(0), 1 – 13.

Miller, J.D., & Kimmel, L.G. (2012). Pathways to a STEM profession. *Peabody Journal of Education*, 87, 26–45

Marsh, H. W., Trautwein, U., Lüdtke, O., Köller, O., & Baumert, J. (2005). Academic self-concept, interest, grades, and standardized test scores: Reciprocal effects models of causal ordering. *Child Development, 76*(2), 397-416.

Morgan, P. L., Farkas, G., Hillemeier, M. M., & Maczuga, S. (2016). Science achievement gaps begin very early, persist, and are largely explained by modifiable factors. *Educational Researcher, 45*(1), 18-35.

Morgan, P. L., Farkas, G., & Wu, Q. (2009). Five-year growth trajectories of kindergarten children with learning difficulties in mathematics. *Journal of Learning Disabilities, 42*(4), 306-321.

Museus, S.D., Palmer, R.T., Davis, R.J., & Maramba, D.C. (2011). Special issue: Racial and ethnic minority students' success in STEM education. *ASHE Higher Education Report*, 36, 1–140

NAEP. (2016). *The nation's report card: Science 2015*. (No. NCES 2012-465). Washington, DC: Institute of Education Sciences, U.S. Department of Education.

National Academies of Sciences, Engineering, and Medicine. (2016). *Science literacy: Concepts, contexts, and consequences.* Washington, DC: The National Academies Press. doi:10.17226/23595.

National Research Council. (2011). *Successful K-12 STEM education: Identifying effective approaches in science, technology, engineering, and mathematics*. Washington, D.C.: National Academies Press.

National Research Council. (1996). *National Science Education Standards*. Washington, DC: The National Academies Press. doi:https://doi.org/10.17226/4962.

National Research Council. (2009). *Learning Science in Informal Environments: People, Places, and Pursuits.* Committee on Learning Science in Informal Environments. Philip Bell, Bruce Lewenstein, Andrew W. Shouse, and Michael A. Federm (Eds.). Board on Science Education, Center for Education. Division of Behavioral and Social Sciences and Education. Washington, DC: The National Academies Press.

National Science Board. (2016). *Science and Engineering Indicators 2016*. Arlington, VA: Natl. Cent. Sci. Eng. Stat. http://www.nsf.gov/statistics/seind16/

NGSS Lead States. (2013). *Next Generation Science Standards: For States, By States*. Washington, DC: The National Academies Press.

No Child Left Behind Act of 2001, Public Law 107-110. (2002). Retrieved from http://www2.ed.gov/policy/elsec/leg/esea02/index.html

O'Reilly, T., & McNamara, D. S. (2007). The impact of science knowledge, reading skill, and reading strategy knowledge on more traditional "high-stakes" measures of high school students' science achievement. *American Educational Research Journal, 44*(1), 161-196.

Peske, H.G. & Haycock, K. (2006). Teaching inequality: How poor and minority students are shortchanged on teacher quality. *The Education Trust*. Retrieved March 16, 2017 from http://files.eric.ed.gov/fulltext/ED494820.pdf

President's Council of Advisors on Science and Technology. (2012). *Engage to excel: Producing one million additional college graduates with degrees in science, technology, engineering, and mathematics*. Washington, DC: PCAST.

http://www.whitehouse.gov/sites/default/files/microsites/ostp/pcast-executive-report-final_2-13-12.pdf. Accessed August 2016.

Phillips, M., Crouse, J., & Ralph, J. (1998). Does the black-white test score gap widen after children enter school? (pp. 229–272), in C. Jencks and M. Phillips (Eds.), *The Black-White Test Score Gap.* Washington, DC: The Brookings Institute.

PISA. (2013). *Highlights from PISA 2012: Performance of U.S. 15-year old students in reading, mathematics, and science literacy in an international context.* (No. NCES 2011-004). Washington, D.C.: National Center for Education Sciences; U.S. Department of Education.

Pomerantz, E. M., Altermatt, E. R., & Saxon, J. L. (2002). Making the grade but feeling distressed: Gender differences in academic performance and internal distress. *Journal of Educational Psychology, 94*, 396–404

Quinn, D. M. (2015). Kindergarten black–white test score gaps: Re-examining the roles of socioeconomic status and school quality with new data. *Sociology of Education*, *88*(2), 120-139.

Quinn, D. M., & Cooc, N. (2015). Science achievement gaps by gender and race/ethnicity in elementary and middle school: Trends and predictors. *Educational Researcher*, *44*(6), 336-346.

Reardon, S. F., & Galindo, C. (2009). The Hispanic-White achievement gap in math and reading in the elementary grades. *American Educational Research Journal*, *46*(3), 853-891.

Reed, D.K., Petscher, Y., & Truckenmiller, A.J. (2016). The contribution of general reading ability to science achievement. *Reading Research Quarterly*.

Riegle-Crumb, C., King, B., Grodsky, E., & Muller, C. (2012). The more things change, the more they stay the same? Prior achievement fails to explain gender inequality in entry into STEM college majors over time. *American Educational Research Journal*, *49*, 1048–73.

Robinson, M. & Ochs, G.T. (2008). Determining why students take more science than required in high school. *Bulletin of Science, Technology & Society*, 28, 338-348.

Rothwell, J. (2013). *The hidden STEM economy*. Washington, DC: Brookings Institute.

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. New York: Wiley.

Sadler, P. M., Sonnert, G., Hazari, Z., & Tai, R. (2012). Stability and volatility of STEM career interest in high school: A gender study. *Science Education*, *96*(3), 411-427.

Schneider, B., Swanson, C.B., Riegle-Crumb, C. (1998). Opportunities for learning: Course sequences and positional advantages. *Social Psychology of Education, 2*: 25–53

Schuelka, M. (2013). Excluding students with disabilities from the culture of achievement: The case of the TIMSS, PIRLS, and PISA. *Journal of Education Policy*, *28*(2), 216 – 230.

Snow, C. E. (2010). Academic language and the challenge of reading for learning about science. *Science*, 328, 450–542.

Stanovich, K.E. (1986) Matthew effects in reading: Some consequences of individual differences in the acquisition of literacy. *Reading Research Quarterly*, *21*, 360-406.

Stevens, J. J., & Schulte, A. C. (2017). The interaction of learning disability status and student demographic characteristics on mathematics growth. *Journal of Learning Disabilities*, *50*(3), 261-274.

Tourangeau, K., Nord, C., Lê, T., Sorongon, A.G., Hagedorn, M.C., Daly, P., & Najarian, M. (2015). *Early Childhood Longitudinal Study, Kindergarten Class of 2010–11 (ECLS-K:2011) User's Manual for the ECLS-K:2011 Kindergarten Data File and Electronic Codebook, Public Version* (NCES 2015-074). U.S. Department of Education. Washington, DC: National Center for Education Statistics.

Tourangeau, K., Nord, C., Lê, T., Wallner-Allen, K., Hagedorn, M.C., Leggitt, J., & Najarian, M. (2015). *Early Childhood Longitudinal Study, Kindergarten Class of 2010–11 (ECLS-K:2011) User's Manual for the ECLS-K:2011 Kindergarten–First Grade Data File and Electronic Codebook, Public Version* (NCES 2015-078). U.S. Department of Education. Washington, DC: National Center for Education Statistics.

Tourangeau, K., Nord, C., Lê, T., Wallner-Allen, K., Vaden-Kiernan, N., Blaker, L. & Najarian, M. (2017). *Early Childhood Longitudinal Study, Kindergarten Class of 2010–11 (ECLS-K:2011) User's Manual for the ECLS-K:2011 Kindergarten–Second Grade Data File and Electronic Codebook, Public Version* (NCES 2017-285). U.S. Department of Education. Washington, DC: National Center for Education Statistics.

U.S. Census Bureau (2012a). *American Community Survey*.

U.S. Census Bureau. (2012b). *Most children younger than age 1 are minorities*. Retrieved from http://www.census.gov/newsroom/releases/archives/population/cb12-90.html. U.S. Department of Education, Institute of Education Sciences

Walberg, H. J., & Tsai, S. L. (1983). Matthew effects in education. *American Educational Research Journal*, *20*(3), 359-373.

Wang, X. (2013). Why students choose STEM majors: motivation, high school learning, and postsecondary context of support. *American Educational Research Journal, 50*(5), 1081–1121.

Weisner, T. S. (1996). The 5 to 7 transition as an ecocultural project. In A. J. Sameroff & M. M. Haith (Eds.), *The Five to Seven Year Shift: The Age of Reason and Responsibility* (pp. 295 – 326). Chicago, IL: University of Chicago.

Wenner, G. (2001). Science and mathematics efficacy beliefs held by practicing and prospective teachers: A 5-year perspective. *Journal of Science Education and Technology*, *10*(2), 181 – 187.

WestEd. (2011). *High hopes – Few opportunities: The status of elementary science education in California*. Santa Cruz, CA: The Center for the Future of Teaching and Learning at WestEd.

White, I. R., Royston, P., & Wood, A. M. (2011). Multiple imputation using chained equations: Issues and guidance for practice. *Statistics in Medicine*, *30*, 377-399. doi: 10.1002/sim.4067

Xie, Y., & Shauman, K. A. (2003). *Women in science career processes and outcomes*. Cambridge, MA: Harvard University Press