The Impact of a Curricular and Professional Development Intervention on Elementary Science Teachers’ Perceptions of High-Stakes Testing and Accountability

Georgina O. Lindskoog

University of Miami, miacanes@hotmail.com

Follow this and additional works at: https://scholarlyrepository.miami.edu/oa_dissertations

Recommended Citation


https://scholarlyrepository.miami.edu/oa_dissertations/1163

This Open access is brought to you for free and open access by the Electronic Theses and Dissertations at Scholarly Repository. It has been accepted for inclusion in Open Access Dissertations by an authorized administrator of Scholarly Repository. For more information, please contact repository.library@miami.edu.
UNIVERSITY OF MIAMI

THE IMPACT OF A CURRICULAR AND PROFESSIONAL DEVELOPMENT INTERVENTION ON ELEMENTARY SCIENCE TEACHERS’ PERCEPTIONS OF HIGH-STAKES TESTING AND ACCOUNTABILITY

By

Georgina O. Lindskoog

A DISSERTATION

Submitted to the Faculty of the University of Miami in partial fulfillment of the requirements for the degree of Doctor of Philosophy

Coral Gables, Florida

May 2014
A dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy

THE IMPACT OF A CURRICULAR AND PROFESSIONAL DEVELOPMENT INTERVENTION ON ELEMENTARY SCIENCE TEACHERS' PERCEPTIONS OF HIGH-STAKES TESTING AND ACCOUNTABILITY

Georgina O. Lindskoog

Approved:

Okhee Lee, Ph.D.
Professor of Teaching and Learning

Walter G. Secada, Ph.D.
Professor of Teaching and Learning

Jennifer M. Langer-Osuna, Ph.D.
Assistant Professor of Teaching and Learning

Soyeon Ahn, Ph.D.
Assistant Professor of Educational and Psychological Studies

Jaime Maerten-Rivera, Ph.D.
Senior Research Associate
Department of Teaching and Learning

M. Brian Blake, Ph.D.
Dean of the Graduate School
High-stakes testing and accountability (HSTA) is a topic largely debated because of the effects it can have on student learning. Teachers’ perceptions help determine teaching practices, ultimately teaching practices impact student learning, and unintended consequences of HSTA have been found to result in teaching practices that do not maximize meaningful student learning. The purpose of this study was to examine elementary science teachers’ perceptions of HSTA prior to an intervention and if participation in the intervention changed their perceptions. In addition, teacher professional variables (i.e., years of teaching and number of science courses taken) and school contextual factors (i.e., percent of Black students, percent of low SES students, and percent of limited English proficient students) were examined as predictors of initial perceptions of HSTA and change in perceptions over time.

The study involved 220 teachers in a large, urban district who taught high numbers of English language learners. Teachers completed questionnaire and background forms. Data were analyzed using longitudinal multilevel modeling and interpreted using social cognitive theory to more completely understand the complexities of human functioning and explain its causality.

Results showed negative perceptions of HSTA and no impact by the intervention on treatment teachers’ perceptions of HSTA compared to control teachers during the
three-year intervention. Themes in the literature that teachers perceive standards positively, HSTA negatively, and the impact of HSTA on teaching practices negatively were supported by results in this study. Teachers’ perceptions of HSTA at baseline were overall consistent across teacher professional variables and school contextual factors tested. At baseline, teachers’ perception that standards provided good guidelines was predicted by years of teaching.
“When you see someone putting on his big boots, you can be pretty sure that an adventure is going to happen.”

A. A. Milne, *Winnie-the-Pooh*

---

**This work is dedicated to G, I, & T**

who inspire me every day to put on my big *brown* boots and experience spectacular adventures!
ACKNOWLEDGEMENTS

An endeavor such as a dissertation study is unlikely to be the result of the effort of just one person. I have been blessed with many individuals who have thoughtfully assisted me. While I cannot list everyone by name, I sincerely hope that each of you who have been a part of my journey know that I recognize and appreciate very much what you have done on my behalf.

Thank you to my professors, classmates, P-SELL colleagues, and the teachers and students at the schools where I have taught and worked. I have found my way by utilizing what I learned from each of you. Special recognition goes to Dr. Okhee Lee, my academic advisor and dissertation director and to the other members of my qualifying exam and dissertation committees: Dr. Walter Secada, Dr. Jennifer Langer-Osuna, Dr. Eduardo Negueruela, Dr. Jaime Maerten-Rivera, and Dr. Soyeon Ahn. Without you this study would not be as rigorous.

For their continued support and encouragement, I am especially appreciative to my family – not only the one I was born into, but also to the members I have been fortunate to add throughout my life. I have no words that rightfully acknowledge the role played by my mother and father not just in this endeavor, but in everything I undertake. Thank you to my sons for motivating me to extend my work in science education. And my eternal gratitude to my husband whose dedication to help me find the focus to overcome the challenges I confronted, and the necessary confidence to tackle such a complex task, allowed me to reach this milestone.

To each of you, ¡mil gracias!
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>LIST OF FIGURES</td>
<td>vi</td>
</tr>
<tr>
<td>LIST OF TABLES</td>
<td>vii</td>
</tr>
<tr>
<td>LIST OF ABBREVIATIONS</td>
<td>viii</td>
</tr>
<tr>
<td>Chapter</td>
<td></td>
</tr>
<tr>
<td>1 INTRODUCTION</td>
<td>1</td>
</tr>
<tr>
<td>2 TEACHERS’ PERCEPTIONS OF HIGH-STAKES TESTING AND ACCOUNTABILITY</td>
<td>7</td>
</tr>
<tr>
<td>3 CONCEPTUAL FRAMEWORK: SOCIAL COGNITIVE THEORY</td>
<td>20</td>
</tr>
<tr>
<td>4 METHOD</td>
<td>33</td>
</tr>
<tr>
<td>5 RESULTS</td>
<td>53</td>
</tr>
<tr>
<td>6 DISCUSSION AND IMPLICATIONS</td>
<td>64</td>
</tr>
<tr>
<td>REFERENCES</td>
<td>76</td>
</tr>
<tr>
<td>Appendix</td>
<td></td>
</tr>
<tr>
<td>A SYNTHETIC REVIEW OF HIGH-STAKES TESTING AND ACCOUNTABILITY IN</td>
<td>82</td>
</tr>
<tr>
<td>ELEMENTARY SCIENCE EDUCATION IN RELATION TO STUDENT DIVERSITY AND</td>
<td></td>
</tr>
<tr>
<td>EQUITY</td>
<td></td>
</tr>
<tr>
<td>B TEACHER QUESTIONNAIRE: HIGH-STAKES TESTING AND ACCOUNTABILITY SECTION</td>
<td>121</td>
</tr>
<tr>
<td>C TEACHER BACKGROUND FORM</td>
<td>124</td>
</tr>
</tbody>
</table>
# LIST OF FIGURES

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>SCT Model of Human Functioning</td>
<td>22</td>
</tr>
<tr>
<td>2.</td>
<td>SCT Model of Science Teachers’ Functioning in a HSTA Context</td>
<td>24</td>
</tr>
<tr>
<td>3.</td>
<td>Impact of the Intervention on Science Teachers’ Perceptions of HSTA</td>
<td>30</td>
</tr>
</tbody>
</table>
### LIST OF TABLES

<table>
<thead>
<tr>
<th>Table</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Number of Participating Schools and Teachers by Year</td>
<td>33</td>
</tr>
<tr>
<td>2. Teacher Demographic and Professional Characteristics</td>
<td>34</td>
</tr>
<tr>
<td>3. Designation of Outcome Variables</td>
<td>39</td>
</tr>
<tr>
<td>4. Intercorrelations Among Teacher Predictor Variables</td>
<td>40</td>
</tr>
<tr>
<td>5. Patterns of Teacher Questionnaire Data Collection</td>
<td>41</td>
</tr>
<tr>
<td>6. Descriptive Statistics for Teachers’ Responses to the Questionnaire</td>
<td>43</td>
</tr>
<tr>
<td>7. Values of Time Variables</td>
<td>47</td>
</tr>
<tr>
<td>8. Results for ITP Scale</td>
<td>54</td>
</tr>
<tr>
<td>9. Results for SGG Ordinal Model (Proportional Odds)</td>
<td>58</td>
</tr>
<tr>
<td>10. Results for APR Ordinal Model (Proportional Odds)</td>
<td>60</td>
</tr>
<tr>
<td>11. Results for ESK Ordinal Model (Proportional Odds)</td>
<td>62</td>
</tr>
</tbody>
</table>
LIST OF ABBREVIATIONS

ELL: English Language Learner
ESE: Exceptional Student Education
ESOL: English for Speakers of Other Languages
HSTA: High-Stakes Testing and Accountability
LEP: Limited English Proficient
MLM: Multilevel modeling
NCLB: No Child Left Behind
PD: Professional Development
SCT: Social Cognitive Theory
SSS: School Site Support
CHAPTER 1

Introduction

Educational researchers have long acknowledged the role of teachers’ perceptions in their teaching practices and the impact on student outcomes (Anderson, 2011; Good & Brophy 1987; Lortie, 1975). Graybill (1997) reported that “attitudes and expectations influence the climate of our classrooms, shaping what we teach and how we teach it, and thus have an impact on the achievement of our students” (p. 311). When Pajares (1992) studied the complexity of examining teachers’ beliefs in educational research, he concluded that examining teachers’ beliefs is necessary and valuable. A strong argument exists that not only do teachers’ perceptions matter in teaching and learning, but that it matters in testing and accountability (Cimbricz, 2002). A review of the literature concluded that:

a relationship between state-mandated testing and teacher beliefs and practice is consistently confirmed. State-mandated tests do matter and do influence what teachers say and do in their classrooms. But while there is overall agreement that a relationship between the two does exist, the nature of that relationship is neither simple nor easy and requires further elucidation. (Cimbricz, 2002, p. 4)

In science education there has been a call in the literature that “a clear understanding of educators’ perceptions of the impacts of current test-based accountability policies should guide the development and implementation of the next generation of national science standards and subsequent large-scale assessments” (Anderson, 2011, p. 104).

Central to science education reform is the idea that “all students deserve and must have the opportunity to become scientifically literate” (National Research Council
This dual goal of promoting high academic achievement for all while ensuring educational equity in science education is arguably laudable. Empirical research supports that inquiry-based teaching and learning is an effective way to attain these goals in science (Akkus, Gunnel, & Hand, 2007). High-stakes testing and accountability (HSTA) has been implemented to ensure that all students receive the same high quality education prescribed by ongoing reform efforts. However, despite promoting such reform-based teaching and learning in science education in a HSTA context, learning gains do not meet the expectations of many education stakeholders and equitable learning opportunities for all in science remain elusive (Lee & Buxton, 2008).

It is reasonable to conclude that the opportunity for all students to receive a meaningful academic experience in science is being complicated, most likely by a variety of key factors. This study posits that one such factor may be the unintended consequences of HSTA. The Board on Testing and Assessment of the Committee on Incentives and Test-Based Accountability in Public Education identified that “incentives will often lead people to find ways to increase measured performance that do not also improve the desired outcomes” (National Research Council [NRC], 2011, p. 2). There is also evidence that in science education:

- When mathematics and English language arts are the focus of accountability, science education was marginalized (Jennings & Rentner, 2006).
- When science is assessed and included in accountability, the multiple-choice format of tests and possible sanctions for low performance led to more didactic teaching practices, countering the preferred inquiry-based method.
reported to result in more meaningful student learning (Coble, 2006; Jenkins, 2000; Settlage & Meadows, 2002).

- A variety of language demands associated with science assessments for English language learners (ELLs) were identified as possibly resulting in test scores that do not accurately reflect ELLs science knowledge (Lyon, Bunch & Shaw, 2012; Shaw, Bunch, & Geaney, 2010).

Consequently, it is hypothesized that an inquiry-based learning intervention designed to enhance teacher capacity and that takes into account student diversity and the HSTA context would ameliorate teachers’ perceptions of HSTA and lead to teachers who would be more likely to promote challenging and equitable learning opportunities for all. Identifying such successful strategies for improving the capacity of teachers to offer instructional support for more meaningful science learning to a diverse student population in a HSTA environment is necessary primarily for three reasons. First, ongoing science education reform promotes the use of inquiry-based learning to achieve a scientifically literate populace. Second, HSTA is well-established in U.S. education as the primary measure of educational reform. Third, U.S. schools are diverse and expected to become more so (Buckley, 2010).

A university in collaboration with a large, urban southeastern U.S. school district, developed a year-long, 5th grade science inquiry-based curriculum. The resulting curricular and professional development intervention was implemented with a group of treatment teachers over a three-year period in a state where science is a part of school accountability and a statewide high-stakes science assessment is administered in fifth grade. The intervention was designed to enhance teacher capacity by providing three
complementary components – curriculum, professional development (PD), and school site support (SSS) – to improve teacher content knowledge and the quality of teaching practices in a HSTA context for teachers teaching a diverse student population that included high numbers of ELLs. Curriculum, PD, and SSS promoted the use of inquiry-based science as an effective method to provide meaningful student learning of state science standards. The intervention was expected to directly impact teacher content knowledge and teaching practices and indirectly to ameliorate teachers’ perceptions of HSTA. The research design included a control group of teachers who did not receive the intervention.

This study was part of the larger evaluation of the impact of the intervention and looked at teachers’ perceptions of HSTA because pre-existing beliefs are key determinants of human behavior. Teachers’ beliefs help determine teaching practices, ultimately teaching practices impact student learning, and unintended consequences of HSTA have been found to result in teaching practices that do not maximize meaningful student learning. The purpose of this study was to examine elementary science teachers’ perceptions of HSTA prior to an intervention and if participation in the intervention changed the teachers’ perceptions of HSTA. In addition, teacher professional variables and school contextual factors were examined as predictors of initial perceptions of HSTA and change of perceptions over time. Specifically, it aimed to answer the following research questions:

1: What were elementary science teachers’ perceptions of HSTA in a large, urban school district prior to any intervention (i.e., at baseline)?
2: Did teacher professional variables (i.e., years of teaching and number of science courses taken) and school contextual factors (i.e., percent of Black students, percent of low SES students, and percent of limited English proficient [LEP] students) predict elementary science teachers’ perceptions of HSTA at baseline?

3: Did elementary science teachers’ perceptions of HSTA in the treatment group change compared to the control group?
   a. Was there a difference in teachers’ perceptions between the treatment and control groups each year (i.e., Year 1, Year 2, and Year 3)?
   b. Was there a trend in differences in teachers’ perceptions between the treatment and control groups over the 3-year period?

4: Did teacher professional variables (i.e., years of teaching and number of science courses taken) and school contextual factors (i.e., percent of Black students, percent of low SES students, and percent of LEP students) predict elementary science teachers’ perceptions of HSTA in the treatment group compared to the control group?
   a. Was there a difference in variations between the treatment and control groups each year (i.e., Year 1, Year 2, and Year 3)?
   b. Was there a trend in variations between the treatment and control groups over the 3-year period?

Identifying the longitudinal impact of the intervention on teachers’ perceptions of HSTA and variations in perceptions is important because it provides empirical evidence of whether or not a well-designed, inquiry-based learning curricular and professional
development intervention can result in perceptions that encourage teachers to be more likely to promote challenging and equitable learning opportunities for all in a HSTA context. Data analysis was conducted using longitudinal multilevel modeling (MLM). Results were interpreted using social cognitive theory (SCT) to more completely understand the complexities of human nature and explain causality. The results of the study identified the effectiveness of the intervention to ameliorate teachers’ perceptions of HSTA over time and how teachers’ perceptions are related to their professional background and school contextual factors.
CHAPTER 2

Teachers’ Perceptions of High-Stakes Testing and Accountability

Defining HSTA Policy

Assessment in the form of standardized testing that is used for the purpose of accountability is referred to as high-stakes testing (Resnick, 2000). Accountability indicates policies that mandate large-scale testing to measure attainment of state and/or federal goals, such as Adequate Yearly Progress (AYP) mandated by No Child Left Behind (NCLB). Accountability systems usually reward schools for high performance and impose sanctions on schools that fail to move students in all identified subgroups toward proficiency. The demands for accountability largely come from politicians and policymakers (Smith & Fey, 2000).

By 2005-2006, NCLB mandated that reading and mathematics be tested annually in grades 3 through 8 and once in high school in all states. Beginning in 2007-2008, NCLB mandated that science be tested at least once in grades 3 through 5, 6 through 9, and 10 through 12, although AYP in science was not required to be measured and reported. Disaggregated results of mandated testing by NCLB for demographic subgroups of students by major ethnic/racial groups, economically disadvantaged students, LEP students, and students with disabilities are made publicly available annually. In some states, science is made part of school accountability at specific grade levels, but generally contributes less to the overall school performance than reading and mathematics.

Whether or not to include student science assessment in accountability measures is debatable. The use of science achievement as a factor of accountability is an option that
is not widespread among states. Judson (2010) reported only 11 states where science contributes appreciably to accountability. This may be explained by various factors such as existing resistance to more high-stakes testing and economic limitations. Nevertheless, standardized testing and test-based accountability policies traditionally used in the areas of English language arts and mathematics are being used in the area of science to determine school grades in some states, such as the state where this study was conducted. And it has been argued that in a well-established context of HSTA the best way to promote science education may be to work through existing testing and accountability structures to avoid marginalization of the subject (Judson, 2013; Ravitch, 2010). Judson (2010) recommended that science assessments be administered and that science achievement results be used in calculations such as AYP.

In the state in which this study took place, development of science test items began in 2000. A field test of these items was conducted with a representative sample in spring 2002 and the first field test was administered in spring 2003. Beginning in 2006-2007, science became a part of school accountability and a statewide high-stakes science assessment is administered in fifth, eighth, and eleventh grades. The elementary schools in the state are evaluated (called the school grade) based on the state test scores in reading, mathematics, writing, and science. In reading and mathematics, performance above mastery and learning gains at third, fourth, and fifth grades count for 3/8 of the annual evaluation, respectively. In writing, performance above mastery at fourth grade counts for 1/8. In science, performance above mastery at fifth grade counts for 1/8.

In 2010-2011, science test scores for ELLs were excluded as part of the school grade during their first two years in English for Speakers of Other Languages (ESOL)
programs, after which science test scores were included regardless of their English proficiency. Beginning in 2011-2012, only ELLs who have less than one year of school time in the U.S. are excluded from counting toward a school’s grade. State law requires that ELLs be provided accommodations when taking the state test, including bilingual dictionaries and additional time to complete the test.

During the intervention, the statewide high-stakes science assessment administered in fifth grade included 55-60 multiple choice items of which 60-70% addressed annually-assessed benchmarks and 30-40% addressed content-sampled benchmarks (i.e., assessed once every 2-3 years). The science test scores were available in terms of scale scores and mastery levels with level 3 and above meeting high standards of *passing*. At the end of Year 1 of the intervention, the state transitioned to a new comprehensive assessment in science to align with the new state science standards. Results were reported using three measures, including scale scores, content area scores, and mastery levels.

In 2010-2011, 51% of the fifth grade students in the state met high standards in science, while 49% of the fifth grade students in the school district where the study took place met high standards in science. In 2011-2012, 51% of the fifth grade students in the state met high standards in science, while 48% of the fifth grade students in the school district met high standards in science. In 2012-2013, 53% of the fifth grade students in the state met high standards in science, while 51% of the fifth grade students in the school district met high standards in science.
Literature Review

This review focused on the literature available on teachers’ perceptions of HSTA in science. Since the student sample in this study was ethnically and linguistically diverse, special consideration was given to including literature on teachers’ perceptions of diverse student populations, especially ELLs who were taught in high numbers in the district where this study took place. For a more comprehensive review of teachers perceptions of HSTA see Appendix A.

Teachers’ perceptions of HSTA. Even before high-stakes testing was the educational norm in the United States, Jackson (1968/1990) reported in his now recognized classic work, *Life in Classrooms*, that elementary teachers showed “a general distrust of tests” (1990, p. 124). Teachers saw a very limited number of useful instruments, believed standardized test results were not reported to the teachers consistently, and that the data arrived too late in the school year to do much good. Teachers further believed that children behaved atypically on tests and that performance on achievement tests was more a reflection of native ability than of teaching effectiveness (Jackson, 1990).

As a standards movement has become the focus of U.S. education, more emphasis has been placed on test-based accountability systems as a means to measure whether all students are achieving proficiency in the standards. NCLB, passed by Congress with bipartisan support and signed into law by President George H. Bush in January 2002, has made HSTA an integral part of every U.S. classroom and generated substantial debate.

A decade after the implementation of NCLB, it has become obvious that its mandate – by 2014 100% of U.S. students would be proficient in reading and
mathematics – would not be achieved. In September 2011, the U.S. Department of Education reported the following to states and districts:

While NCLB helped State and local educational agencies (SEAs and LEAs) shine a bright light on the achievement gap and increased accountability for groups of high-need students, it inadvertently encouraged some States to set low academic standards, failed to recognize or reward growth in student learning, and did little to elevate the teaching profession or recognize the most effective teachers. (Duncan, 2011)

As a result, a process was put in place for states and districts to request flexibility regarding specific requirements of NCLB. By August 2012, 32 states and the District of Columbia had been granted waivers. To be eligible to receive the waiver, states had to develop and implement plans to improve educational outcomes as outlined by the Obama education agenda, such as using students’ standardized test scores as part of their teacher evaluation systems. The state in which this study took place applied for and was granted a waiver from NCLB in February 2012.

Although NCLB has largely been criticized, positive impact has also been noted. First, studies reported that student performance on high-stakes testing has improved (Hamilton et al., 2007; Rentner et al., 2006). For example, Massachusetts saw average student scores on national tests shift to be the top in the nation in the years following the institution of high-stakes testing (Reville, 2004). Bishop (2000) found that curriculum-based external exit exam systems, such as those used in New York and Canada, resulted in higher student performance compared to students not required to take exit exams. Second, non-dominant groups are now part of the educational discourse (Cizek, 2001).
For example, in North Carolina fourth and fifth grade students with learning disabilities received more attention since their inclusion in high-stakes testing, leading to a marked increase in reading ability (Schulte, Villwock, Whichard, & Stallings, 2001). Third, NCLB has led to increased instruction time for subject areas that previously were underemphasized, such as science, in years when high-stakes testing is administered (Rodgers, 2006).

As HSTA policy in science has become more common, studies have focused on examining science teachers’ perceptions. Anderson (2011), in his synthesis of studies on science education and test-based accountability policies, identified 35 studies of which the majority used ethnographic or case study methodology. These studies included work conducted in 13 elementary schools, 9 middle schools, and 13 high schools from across the United States, in addition to three studies conducted in England, Canada, and England and Wales. From these studies, some broadly held science teachers’ perceptions of HSTA for science education have emerged. For example, science teachers perceived standards positively (Anderson, 2011), but believed that state science tests did not accurately represent student science knowledge (Font-Rivera, 2003). Diamond and Spillane (2004) noted that in schools on probation for not having met achievement standards in reading and mathematics, the instructional time dedicated to science was decreased or even eliminated. However, science teachers of grade levels where high-stakes testing was not administered reported minimal impact on science instructional time (Wideen, O’Shea, Pye, & Ivany, 2007). Other studies found that science teachers saw the focus in science education turned to students near the cut score for proficiency on state science tests at the expense of high-achieving students (Hamilton et al., 2007).
Science teaching practices. Science teachers credited high-stakes testing with increasing curricular rigor and better aligning curriculum with standards and assessments (Anderson, 2011; Hamilton et al., 2007). Vogler (2002) found that a stratified random sample of teachers teaching at least one section of tenth grade English, mathematics, or science in a public high school in Massachusetts self-reported a greatly increased use of inquiry-based lessons when high-stakes test scores first became public. More specifically, 25% of teachers reported increased use of lab equipment, 31% reported increased use of modeling, and 42% reported increased use of computers. Additionally, 12% of teachers reported decreased use of textbook-based assignments and 36% reported decreased lectures. Data in this study were obtained from a 54-question survey analyzed using frequency tables, cross tabulations, and grouping of comments according to content. Vogler (2002) does not provide details as to how strategies were implemented, only whether or not the teachers felt they used them. The teachers most likely to change their teaching practices were those who had been teaching for 13-19 years, while both more and less experienced teachers reported less change. The most common reason teachers gave for changing their teaching practices was to help students pass the test so the students could graduate. Barksdale-Ladd and Thomas (2000) found that “no teacher argued for a lack of standards or for dismissing assessment” (p. 395).

Contradictory to the results presented by Vogler (2002), other studies found that teachers reported a decrease in the use of inquiry lessons and an increase in fact-based learning (Coble, 2006; Jenkins, 2000; Settlage & Meadows, 2002). Science teachers also believed that teachers of non-tested courses used more inquiry-based curricula than teachers of courses that are being assessed as part of accountability policies (Coble,
Moreover, science teachers believed that high-stakes testing led to a focus on teaching to the test (Font-Rivera, 2003; Hamilton et al., 2007) and that science teachers devoted significant class time to teaching test-taking skills (Font-Rivera, 2003; Hamilton et al., 2007).

**Science teachers’ sense of professional well-being.** Science teachers felt less professional satisfaction as a result of HSTA (Anderson, 2011; Jenkins 2000). They perceived that HSTA policies resulted in a decreased sense of autonomy (Mintrop, 2004), an increased level of stress and decreased morale (Aronson, 2007; Rentner et al., 2006), and pressure to increase test scores (Font-Rivera, 2003). In addition, they agreed that high-stakes testing led them to teach in ways that compromised their view of effective teaching (Aronson, 2007) or did not allow them to explore topics of student interest (Galton, 2002).

**Science teachers’ perceptions of the equitable teaching of ELLs.** One of the most significant outcomes of HSTA has been the increased focus on achievement gaps and the learning needs of marginalized student groups. Issues of multicultural education are more salient as the U.S. student population grows more diverse. According to the U.S. Census, it is projected that racial and ethnic minorities will make up half of all individuals under 19 years of age by 2022 (U.S. Census Bureau, 2012). One of the most significant demographic shifts currently affecting U.S. education is the growing number of ELLs (National Center for Education Statistics, 2012). Among children between 5-17 years old, 21% spoke a language other than English as their primary language at home and spoke English with difficulty (National Center for Education Statistics, 2012). In 2009 over 5 million, or 11% of U.S. public school students in K-12, were identified as
ELLs (National Clearinghouse for English Language Acquisition, 2011). Addressing the needs of ELLs is important as a greater number of teachers are tasked with preparing ELLs for successful performance on high-stakes testing.

Under existing accountability policies, ELLs are expected to take part in all assessments and have their scores count as part of various accountability measures with minimal exceptions. Allowances from assessments and accountability are generally granted for the lowest levels of LEP students for the first two years or in some cases only one year, of full-time attendance in a U.S. school. State laws require that ELLs be provided accommodations when taking high-stakes tests. Accommodations include a variety of adjustments in timing and scheduling of tests, the location where the test is administered, the use of bilingual dictionaries, and the presentation of test materials or test directions. Nevertheless, the validity of inferences for ELLs based on assessment scores has come into question. Keiffer, Lesaux, Rivera, and Francis (2009) found that “accommodations to reduce the impact of limited language proficiency on academic skill assessment are not particularly effective” (p. 1168). The stakes are high not only for schools and teachers, but for individual students as well. Students’ scores are routinely used in determining their placement in advanced courses, as well as promotion or graduation.

The literature is consistent that mainstream teachers need specialized knowledge and preparation to work with ELLs (Bunch, 2013; Lucas, 2011). Many elementary science teachers, however, felt unprepared to teach science (Banilower, Smith, Weiss, Malzahn, Campbell, & Weis, 2013) and to address student diversity (National Center for Education Statistics, 2012). Many teachers believed that the standards implemented were
often not appropriate for the diverse students they were assigned to teach (Barksdale-Ladd & Thomas, 2000). Bilingual education teachers believed that largely because of high-stakes testing they were ignored when it came to contributing to the best course of action to teach ELLs (Palmer & Garcia, 2000). As ELLs are pressured to learn English, the language of the test, there is concern that the students will fall further behind in academic subjects such as science without time to incorporate the scaffolding provided by their native language.

Science teachers viewed HSTA as making it especially difficult to offer equitable learning opportunities for non-dominant student groups, such as ELLs (Penfield & Lee, 2010). Teachers were concerned that students pulled out of content courses for ESOL classes were not receiving the science instruction they were to be tested on (Shaver, Cuevas, Lee, & Avalos, 2007). Wright and Choi (2006) found that teachers viewed state-mandated tests as not being equitable for ELLs:

In terms of accountability, while the overwhelming majority of teachers support the general principle, they believe that high-stakes tests are inappropriate for ELLs and participants provided evidence that the focus on testing is leading to instruction practices for ELLs which fail to meet their unique linguistic and academic needs. (p. 2)

**Impact of interventions on teachers’ perceptions of HSTA.** No published journal articles were found that examined the impact of an intervention on teachers’ perceptions of HSTA. Herring (2009) in his doctoral dissertation assessed the impact of a science outreach program on student achievement, teacher efficacy, and teacher perception. Specifically, one of his research questions addressed whether being involved
in the intervention affected the participants’ perspectives on HSTA. The study employed phenomenological qualitative method using purposeful sampling. Four interviewees responded to a total of 12 semi-structured items including two that were directly related to the accountability issue. Herring (2009) reports that 3 of the 4 participants believed that the intervention “enhanced their students’ ability to perform at a higher level on high stakes tests such as the graduation exam, AP exam, or Criterion Referenced Test (CRT)” (p. 130).

**Teachers’ perceptions by professional variables and school contextual factors.** Few studies examining how teachers’ perceptions vary by professional variables were found. Smith and Kovacs (2011) focused on the teacher variable, years of experience. This study noted that teachers with more years of teaching experience felt more strongly about the negative effects associated with NCLB policy. Stuart Hammer (2004) examined accountability and standards-based reform in middle school science in the year after the state implemented high-stakes testing. Specifically, the study examined whether science teachers’ attitudes about accountability and standards-based reform varied by years of teaching at their current school, teachers’ attitudes toward school administration, percentage of students on free and reduced price lunch, school community environment (urban, suburban, or rural), percentage of minority students at a school, and grade level taught. The results indicated that “only the degree of support teachers receive from the administrators at their school had a consistent influence on the six factors of school reform examined” (p. 94). More support by administrators was associated with more positive teacher attitudes.
Limited doctoral dissertation work included Lowry (2010) who found teachers in high-stakes testing environments spent more time on test preparation than teachers in non-high-stakes environments and that this did not vary by gender or tenure. Mowers’ (2010) dissertation study examined if teachers under the consistent stress of accountability and NCLB showed signs of burnout and whether these signs varied by gender, teacher education, and years of teaching experience. This study noted that “female teachers exhibit slightly more emotional exhaustion and even less personal accomplishment than male teachers and teachers with the most education exhibited the highest level of depersonalization toward their students” (p. 53). Anderson (2012) in his doctoral study examined how state policy influenced science teachers’ perceptions and how these varied by grade level taught, years of experience in science teaching, background in science and science teaching, and teachers’ perceptions of district leadership and policies. There generally was no variation other than “where teachers saw significant district influence they also reported more significant influence of state standards and testing on their teaching. They did not, however, see increased district influence as pushing them away from or toward inquiry or lab-based instruction” (p. 153).

There is substantial research on teachers’ interactions and attitudes with nondominant student populations. These studies found that teachers have cultural biases toward low SES students (Payne, 1994), expectations and speech vary with students’ ethnic background (Tenenbaum & Ruck, 2007), and teachers have different educational aspirations for students based on students’ ethnicity/race (Cheng & Starks, 2002). Smith and Kovacs (2011) noted that teachers in low SES or high racial minority schools felt
more strongly about the negative effects associated with NCLB policy. Teachers believed that nondominant students have more academic difficulties, fewer adaptive behaviors in school, and poorer educational prognoses than students from dominant groups.

The scholarly work that was reviewed appeared in peer reviewed journals or was an accepted dissertation that included empirical data. Few studies, however, used stratified random samples of teachers. Anderson’s 2011 synthesis of studies on science education and test-based accountability policies showed that after removing three studies conducted outside the United States, 63% of the research reviewed was undertaken using an ethnographic or case study methodology that used observations and interviews including open-ended or semi-structured items (Anderson, 2011, Table 1, pp. 4-11). Some studies included very few numbers of schools and/or teachers in their sample. This review also highlighted that existing work in teachers’ perceptions of HSTA has been done nearly twice as often with middle and high school teachers as with elementary teachers.
CHAPTER 3

Conceptual Framework: Social Cognitive Theory

This study focuses on identifying and explaining why teachers’ perceptions of HSTA may exist and how the impact of the intervention may have resulted. In order to examine teachers’ perceptions, SCT was used to help better understand the complexities of human nature and functioning. SCT highlighted what were likely factors influencing teachers’ perceptions, how these factors interacted, and how the factors explain the impact of the intervention. This chapter presents an overview of SCT and then how a SCT lens offers a view of the science teacher in a HSTA context.

Overview of SCT

SCT stems from the theory of social learning and imitation proposed by Miller and Dollard in 1941 that highlighted drive reduction principles. Drive reduction proposed that drives were stimuli that produce discomfort and impel individuals to act when one encounters a cue that triggers a behavior that could reduce the drive. When the behavior succeeds in reducing the drive (i.e., the discomfort), the behavior or pattern of behaviors is likely to result in repetition of the behavior or pattern. This theory was expanded by Bandura and Walters in 1963 to include observational learning and vicarious reinforcement. Bandura in 1977 added the concept of self-efficacy. In 1986, in *Social Foundations of Thought and Action: A Social Cognitive Theory*, Bandura put forth a view of human functioning where individuals are proactive, instead of reactive, participants in the development of one’s behavioral patterns. In other words, people are not solely driven by impulses and or environmental pressures; they are self-organizing, self-reflecting, and self-regulating.
According to SCT, human functioning is the result of the bi-directional product of a dynamic interplay of three factors – behavioral, environmental, and personal (Bandura, 1986; Pajares, 1996 & 2002). This triadic reciprocality refers to the mutual action between causal factors. Since individuals are seen to behave with purpose, they do not simply react as a result of past experiences. While past behavior informs, it is the human capacity for forethought that is possibly a more significant influence in future behavior by allowing individuals to anticipate the likely consequences of their prospective actions. “Through exercise of forethought, people motivate themselves and guide their actions anticipatorily” (Bandura, 1986, p.19). When a teacher perceives that the time necessary to conduct inquiry-based learning may be more than what he/she can invest and also meet the demands of HSTA, future behavior may result in less use of inquiry-based science.

Environmental factors influence an individual’s behavior by providing models for behavior. For example, teacher education programs, PD, and reform documents such as the Next Generation Science Standards (NGSS) promote the use of inquiry in the classroom. Personal factors include cognition, affect, and biological components. The process of acquiring knowledge and understanding through thought, experience, and the senses play a significant role in people’s ability to construct reality, self-regulate, encode information, and perform behaviors. Teachers who have improved content knowledge in science would be more confident in their abilities and more likely to use more challenging classroom practices such as inquiry-based science. Pre-existing beliefs (beliefs and perceptions are interpreted in this study to have similar meanings) are key determinants of human functioning. An individual’s beliefs inform and alter their environments and behaviors which, in turn inform and alter existing beliefs. When a
teacher believes that inquiry science leads to more meaningful student learning such teaching practices would be used and opportunities to learn more about such teaching practices would be sought (see Figure 1).

Figure 1. *SCT model of human functioning*


SCT recognizes the impact of environmental factors and biological components such as genetics on human learning, development, behavior, and adaptation and argues that “a theory that denies that thoughts can regulate actions does not lend itself readily to the explanation of complex human behavior” (Bandura, 1986, p. 15). Human agency, individual and collective since individuals in a social context interact with a myriad of individuals who provide opportunities for observational learning, are causal factors of behavior according to SCT. Individuals are active participants in their development who make things happen by their actions. Self-beliefs allow people to control their thoughts, feelings and actions because "what people think, believe, and feel affects how they behave" (Bandura, 1986, p. 25). Individual change also depends on the interaction of the three factors: behavioral, environmental, and personal. It is reasonable to conclude that
teachers who are routinely tasked to prepare students for state mandated tests tied to high-stakes whose outcomes directly impact them, their workplace, and their students would have strong perceptions of HSTA that would likely influence their behavior.

**Science Teachers in a HSTA Context**

In the section that follows, this study draws on each of the three aspects of the SCT triangle to identify the factors and interactions that affect science teachers in a HSTA context. If one views the teacher participants through the social cognitive lens, a teacher’s behavior will be influenced by the specific policy and characteristics of the institutional environment and school organization (i.e., environmental factors), the teaching practices being used (i.e., behavioral factors), and their preexisting perceptions (i.e., personal factors).

According to SCT, reciprocality occurs not only between causal factors but also within the three realms of behavioral, environmental, and personal factors. Whether the relationships are actual or perceived can make little difference on the effect: “False beliefs activate avoidant behavior that keeps individuals out of touch with prevailing reality, thus creating a strong reciprocal interaction between beliefs and action, which is protected from corrective environmental influence” (Bandura, 1986, p. 24). Reciprocal processes likely operating within each of the three SCT realms, as well as the interactions of causal factors, are described in the following sections. See Figure 2 for the dynamics of teacher functioning in a HSTA context.
Figure 2. SCT Model of Science Teachers’ Functioning in a HSTA Context

**Behavioral factors.** In the behavioral realm, actions are related such that their occurrences co-vary positively and negatively. Using inquiry can co-vary positively with more meaningful learning. This relationship is likely to result in the use of inquiry-based learning. Teaching to the test can co-vary positively with higher test scores. This relationship is likely to encourage the use of non-preferred didactic teaching practices. Higher test scores and meaningful student learning can be viewed to co-vary negatively. This interaction is likely to result in internal turmoil that leads to negative perceptions of HSTA.

Teaching practices will influence and be influenced by teachers’ perceptions and HSTA and reform-based science. Research on teachers’ perceptions showed that science teachers reported a decrease in the use of inquiry lessons and an increase in fact-based
learning as a result of HSTA (Coble, 2006; Jenkins, 2000; Settlage & Meadows, 2002). They also believed that teachers of non-tested courses used more inquiry-based curricula than teachers of courses that are being assessed as part of accountability policies (Coble, 2006). High-stakes testing was seen to lead to a focus on teaching to the test (Font-Rivera, 2003; Hamilton et al., 2007) and science teachers reported devoting significant class time to teaching test-taking skills (Font-Rivera, 2003; Hamilton et al., 2007). Plank and Condliffe (2013) examined the influence of HSTA on classroom quality and found that while teachers offered moderate to high levels of emotional support and classroom organization, teachers offered low levels of instructional support, higher-order thinking, feedback between teachers and students about concepts and thought processes, and language modeling. They also reported that “classroom quality is lower when classrooms are under greatest pressure to increase test performance” (Plank & Condliffe, 2013, p. 1152). An emerging body of research on the impact of HSTA on teaching practices shows that accountability influences teaching practices (Hannaway & Hamilton, 2008).

**Environmental factors.** In the environmental realm, factors also interact (e.g., reform-based science and HSTA). The influence of reform-based science is likely to encourage positive perceptions of standards and the use of inquiry compared to HSTA demands which may not. Contradictory messages and perceptions that HSTA make pursuing a preferred behavior, such as delving into topics of student interest, impossible are likely to result in internal turmoil. This in turn is likely to result in teachers feeling a lack of autonomy and decreased sense of professional well-being.

SCT argues that HSTA and reform-based science will in turn influence and be influenced by teachers’ perceptions and teaching practices. An example of HSTA
influencing behavior was recorded by Abrams, Pedulla, and Madaus (2003) who found that teachers requested to be reassigned from tested grades to non-tested grades. Reform-based science education focuses on student inputs and emphasizes inquiry-based learning, critical thinking, evidence-based argumentation, and collaborative problem solving. Reform documents such as the NGSS allocate significant decision-making to teachers, “teachers should rely on quality instructional products and their own professional judgment as the best way to implement the NGSS in classrooms” (NGSS, 2013, XVIII).

As the impact of the reforms is measured by HSTA, pressures to maximize scores in high-stakes assessments, especially in English language arts and mathematics, to meet the requirements of NCLB and state accountability policies are influencing science teaching and learning. Instructional time dedicated to science was decreased or even eliminated in schools on probation for not having met achievement standards in reading and mathematics (Diamond & Spillane, 2004). HSTA focuses on student outcomes, specifically test scores. It comes from the top down and results from decisions made by policymakers who generally do not consult those impacted (Smith & Fey, 2000). Consequences are felt by schools and students, and increasingly are being directed at classroom teachers (National Research Council [NRC], 2011). Teacher expectations as a result of HSTA are very different from the expectations posed by reform-based science. Teachers may feel they are faced with a conundrum of whether to choose teaching practices that promote the goals of science education reform or HSTA.

In the state where this study took place, state science standards guide science instruction. Students take state tests to assess their mastery of standards that have high-
stakes attached, such as helping determine placement in advanced instruction and promotion at third grade. Elementary schools are graded on eight assessment based measures of achievement on current year test scores and components that measure progress from the previous year. State test scores in science in 5th grade are included as part of the accountability policy and count for 1/8 of the total school grade. ELLs, which are in large numbers in this district, are excluded in computations of the school grade with minimal exceptions regardless of their English proficiency. State standards, high-stakes testing, accountability policy, and percentages of ELLs are expected to influence teachers’ perceptions of HSTA.

Personal factors. In the personal realm, processes which reinforce each other can escalate thoughts, feelings, or perceptions. Teachers’ perceptions such as the state accountability policy is not reasonable, state tests are not accurate measures of ELLs science knowledge, and HSTA impacts teaching practices negatively reinforce each other and are likely to escalate negative perceptions of HSTA.

Several themes identified in the literature showed that teachers viewed standards positively, but HSTA negatively (Abrams, Pedulla, & Madaus, 2003; Barksdale-Ladd & Thomas, 2000). They perceived that HSTA affects teaching practices negatively (Aronson, 2007; Coble, 2006; Hamilton, et al. 2007) and that HSTA decreases teachers’ sense of professional well-being (Aronson, 2007; Barksdale-Ladd & Thomas, 2000; Mintrop, 2004). There is abundant research on teachers’ interactions and attitudes with students from non-dominant groups. Some research demonstrates teacher beliefs that nondominant students have more academic difficulties, fewer adaptive behaviors in academic settings, and poorer educational prognoses than students from dominant groups.
Teachers with higher percentages of nondominant students, including ELLs and low SES students, could believe that HSTA is an inadequate measure of their students’ knowledge and thus more unfair than teachers teaching students from dominant groups or higher SES.

Teaching experience and teacher content knowledge could also impact teachers’ perceptions of HSTA. More experienced teachers who taught before the implementation of HSTA were found to have stronger perceptions about HSTA than less experienced teachers who have taught, and possibly also been taught in a HSTA context (Smith and Kovacs, 2011). Teachers who possess greater content knowledge are expected to be more confident in their teaching efficacy. Teachers with improved teacher content knowledge may feel more comfortable implementing inquiry and more confident of its results, thus feeling less challenged by HSTA than teachers having less teacher content knowledge. Banilower et al. (2013) found that many elementary science teachers reported feeling unprepared to teach science.

It was expected that the intervention would moderate existing tensions and help resolve conundrums posed by reform-based science and HSTA and therefore result in a change of teachers’ perceptions of HSTA for treatment teachers. After the intervention, it was anticipated that treatment teachers would be more critical of teaching to the test and more supportive of reform-based science compared to teachers in the control group. This difference would result because on one hand the control group received no resources to address the challenges and opportunities posed by reform-based science and HSTA. On the other hand, the treatment teachers received a variety of support to address the challenges and opportunities in the form of curriculum, PD, and SSS.
**Impact of the intervention.** The bidirectional nature of the factors in SCT suggests that strategies for improvement can be directed at any of the three factors: behavioral, environmental, or personal. The intervention examined in this study was introduced as an environmental factor to improve personal factors (i.e., teacher content knowledge and teachers’ perceptions) and a behavioral factor (i.e., teaching practices). According to SCT, “triadic factors do not operate simultaneously as a wholistic entity, it is possible to gain some understanding of how different segments of two-way causation operate” (Bandura, 1986, p. 25). This study focused on examining the interactions between teachers’ perceptions of HSTA and the intervention. According to Bandura, studying a two-way segment such as in this study between a personal factor (i.e., teachers perceptions) and an environmental factor (i.e., the intervention) is a significant contribution because “clarifying how the various subsystems function interactively advances understanding of how the superordinate system operates” (Bandura, 1986, p. 25). Interpretation of results will serve not only to highlight the impact of the intervention on existing teachers’ perceptions of HSTA, but also explain how teachers’ perceptions may interact with other factors and result in patterns of teaching practices for elementary science teachers (see Figure 3).
Figure 3. *Impact of the intervention on elementary science teachers’ perceptions of HSTA*

The three components of this curricular and professional development intervention working together was anticipated to significantly moderate tensions between the demands posed by reform-based science and those posed by HSTA (actual or perceived). Treatment teachers with an improved teacher content knowledge and possessing resources to facilitate implementation of inquiry science for all the topics included in the state standards and with strategies for ELLs were expected to have more positive perceptions of standards and less negative perceptions of HSTA after the intervention compared to the control teachers.

*Curriculum.* A year-long 5th grade science curriculum that addressed all of the benchmarks in the state science standards and highlighted inquiry-based learning to maximize understanding of science concepts was developed and provided to all students
of teachers’ in the treatment group. A teacher guide that offered strategies, for all students and specifically for ELLs, to promote science inquiry and understanding was developed and provided to all treatment teachers. Supplementary curriculum materials were developed and made available via CD and on-line at the project’s website in response to teachers’ requests.

The student book and teacher guide were revised to reflect new science standards adopted by the state at the conclusion of Year 1. Changes included (1) inclusion of new science standards adopted by the state, (2) the order in which the content was organized, and (3) modification of some inquiry and hands-on activities in response to teacher feedback to increase ease of implementation. Control teachers continued using the same previously adopted text book.

**Professional development.** During the three-year intervention, workshops for treatment teachers were organized to improve teacher content knowledge and teaching practices. At first, PD focused on familiarizing teachers with the curriculum, demonstrating teaching and learning strategies, and engaging teachers in participation of inquiry-based learning and hands-on activities, as well as highlighting the curriculum’s alignment to the state science standards and high-stakes assessment. PD then focused on highlighting changes to the new state science standards and the new high-stakes science assessment aligned with the new standards. As implementation progressed, workshop activities were developed that highlighted inquiry strategies for how to use opportunities for more open-ended questioning in science instruction, recognize what is a good student response, utilize effective ways of grading/providing constructive feedback to students, and develop rubrics for inquiry/open-ended questioning. Finally, a workshop focused on
helping teachers implement student-centered inquiry and language development strategies for ELLs.

**School site support.** Staff provided treatment teachers with individual support during all three years of the intervention. Schools were generally visited every 4 to 6 weeks or more often if a school or teacher requested assistance or if the project staff or the school district administrators identified a school or teacher to be high-needs. Support included planning, teaching/modeling, co-teaching, reflection, delivering supplies, and providing additional resources as requested by teachers for all students, including ELLs (e.g., home-learning suggestions, lab management strategies, and supplemental assessment materials). This one on one provided by SSS especially recognized and addressed individual teacher concerns and in cooperation with the other two components of the intervention were expected to ameliorate perceptions such as HSTA resulted in increased didactic instruction and the state test is not an accurate measure of ELLs science knowledge.

Educational research has argued that teacher perceptions are slow to change. SCT recognizes that “it takes time for a causal factor to exert its influence and interacting factors work their mutual effects sequentially over variable time courses” (Bandura, 1986, p. 25). This study collected data over three years to achieve measurable results and performed a longitudinal analysis. Additional details of the intervention and specifics on how it was implemented are described in the following chapter.
CHAPTER 4

Method

Research Setting and Design

The intervention was implemented in a large, urban school district in the Southeast United States during three consecutive school years: 2010-2011 (Year 1), 2011-2012 (Year 2), and 2012-2013 (Year 3). This research employed a cluster randomized control trial. Of the 238 elementary schools, the district provided a pool of 206 elementary schools after excluding 32 schools from participation, including 23 because they were under close monitoring by the school district due to critically low performance and an additional 9 schools because they had participated in our previous project. From the pool of 206 schools, 64 schools were randomly selected and then randomly assigned to one of two groups: the treatment group \((n = 32)\) and the control group \((n = 32)\). An invitation was sent and all 64 schools agreed to participate. As a school-wide initiative, all fifth grade teachers from these 64 schools agreed to participate.

At the start of Year 1, two control schools elected not to participate in the study. At the beginning of Year 2, one treatment school that participated in Year 1 of the study chose not to continue. There was no change in the number of schools participating in Year 3. See Table 1 for a summary of the number of participating schools and teachers by year of implementation of the intervention.

<table>
<thead>
<tr>
<th>Year</th>
<th>Schools</th>
<th>Total</th>
<th>Teachers</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Treatment</td>
<td>Control</td>
<td></td>
<td>Treatment</td>
</tr>
<tr>
<td>1</td>
<td>32</td>
<td>30</td>
<td>62</td>
<td>122</td>
</tr>
<tr>
<td>2</td>
<td>31</td>
<td>30</td>
<td>61</td>
<td>123</td>
</tr>
<tr>
<td>3</td>
<td>31</td>
<td>30</td>
<td>61</td>
<td>114</td>
</tr>
</tbody>
</table>
Teacher and Student Sample

Teacher sample. Table 2 presents teacher demographic and professional characteristics as reported by the participants. The number of years teaching ranged from 0 to 42 with a mean of 12.09 years ($SD = 8.49$). The number of years teaching at the current school ranged from 0 to 36 years with a mean of 6.77 years ($SD = 6.32$). The number of science courses taken ranged from 0 to 24 with a mean of 4.69 courses taken ($SD = 4.53$). Almost all the teachers (92%) were ESOL endorsed.

Table 2
Teacher Demographic and Professional Characteristics ($N = 359$)

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>Treatment $(n = 183)$</th>
<th>Control $(n = 176)$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Percent (%)</td>
<td>Percent (%)</td>
</tr>
<tr>
<td>Gender</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>18</td>
<td>13</td>
</tr>
<tr>
<td>Female</td>
<td>82</td>
<td>84</td>
</tr>
<tr>
<td>Missing</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>Ethnicity</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hispanic</td>
<td>55</td>
<td>54</td>
</tr>
<tr>
<td>Black non-Hispanic</td>
<td>23</td>
<td>19</td>
</tr>
<tr>
<td>White non-Hispanic</td>
<td>19</td>
<td>19</td>
</tr>
<tr>
<td>Haitian</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>Other</td>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td>Missing</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td>Native Language(s)*</td>
<td></td>
<td></td>
</tr>
<tr>
<td>English</td>
<td>83</td>
<td>79</td>
</tr>
<tr>
<td>Spanish</td>
<td>44</td>
<td>53</td>
</tr>
<tr>
<td>Haitian Creole</td>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>French</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>Other</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>Degrees</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bachelor's</td>
<td>49</td>
<td>47</td>
</tr>
<tr>
<td>Master's</td>
<td>44</td>
<td>42</td>
</tr>
<tr>
<td>Specialist</td>
<td>4</td>
<td>6</td>
</tr>
<tr>
<td>Doctorate</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Missing</td>
<td>2</td>
<td>5</td>
</tr>
<tr>
<td>ESOL Endorsed</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>92</td>
<td>90</td>
</tr>
<tr>
<td>No</td>
<td>8</td>
<td>10</td>
</tr>
</tbody>
</table>

*Note. Multiple native languages could be selected.
**Student sample.** Because this intervention was designed to improve teachers’ content knowledge and teaching practices for diverse student groups, including ELLs, it is anticipated that teachers’ perceptions of HSTA would change as they worked with diverse student groups over the course of the three-year intervention. At baseline the ethnic makeup of the student population in the school district was approximately 65% Hispanic, 24% Black non-Hispanic (including Haitian and Caribbean Islanders), 9% White Non-Hispanic, and 2% who self-identified as Other. Across the school district, 72% of elementary students received free or reduced price lunch programs, 19% participated in ESOL programs, and 11% participated in exceptional student education (ESE) programs. There was no appreciable change in percentages of student groups in the district during the implementation of the intervention.

**The Intervention**

Curriculum, PD, and SSS were designed to complement one another with the goal of improving teachers’ content knowledge and teaching practices in science according to state science content standards for diverse student groups, including ELLs, in a HSTA policy context. The three intervention components are explained below. Throughout the three years of the intervention, the control schools implemented the district adopted science curriculum in *business as usual*.

**Curricular component.** The curricular component was a project-developed year-long 5th grade science curriculum that addressed all of the benchmarks in the state science standards. The intervention provided student books, teacher guides, and supplies to the treatment schools. Class sets of the student book were provided each year and consumable supplies were replenished at the start of subsequent school years. Within the
The student book and teacher guide were revised to reflect new science standards adopted by the state at the conclusion of Year 1. While the design and inquiry orientation of the curriculum remained intact, changes included (1) inclusion of new science standards adopted by the state, (2) the order in which the content was organized, and (3) modification of some inquiry and hands-on activities in response to teacher feedback to increase ease of implementation. Additionally, supplementary curriculum materials were developed and made available via CD and on-line at the project’s website.

**PD component.** During Year 1, treatment teachers were provided with five days of workshops, including 3 days in the summer shortly before school started, 1 day in January, and 1 day in May. Teachers received stipends for attending the summer workshops and schools received substitute payments during school days. The workshops focused on familiarizing teachers with the curriculum, demonstrating strategies, and engaging teachers in inquiry-based learning and hands-on activities, as well as highlighting the curriculum’s alignment to the state science standards and high-stakes assessment.

During Year 2, treatment teachers were provided with another five days of workshops that followed the same schedule as in Year 1 – three days in the summer, 1 day in January, and 1 day in May. At the summer workshops, teachers new to the intervention were separated from returning teachers to meet their needs of familiarizing
with the curriculum, as was the case with the teachers during Year 1. With returning teachers, the primary focus was on highlighting changes to the new state science standards and the new high-stakes science assessment aligned with the new standards. At the January workshop, both the new and returning teachers convened to become familiar with the curriculum that was revised to align with the new science state standards and high-stakes science assessment. The May workshop with both the new and returning teachers focused on unpacking science inquiry – highlighting opportunities for more open-ended questioning in science instruction, making teachers more comfortable and willing to implement inquiry, informing teachers of what is a good student response, providing examples of how to grade/provide constructive feedback to students, and facilitating the development and use of rubrics for inquiry/open-ended questioning.

During Year 3, returning teachers were provided with one additional day of workshop in the summer. The focus was on providing activities to help teachers implement student-centered inquiry and language development strategies for ELLs. Teachers new to the intervention were offered three additional days of workshops in the summer to familiarize them with the curriculum, demonstrate strategies, engage in inquiry-based learning and hands-on activities, and highlight the curriculum’s alignment to the state science standards and high-stakes assessment. At the final meeting in May, both new and returning teachers convened to offer feedback on the intervention and reorient to the district adopted curriculum which they would be transitioning back to the following school year.

**SSS component.** Members of the research team were assigned to provide teachers with support and ensure fidelity of implementation during all three years of the
intervention. Three members of the research team were assigned in Year 1 and Year 2 and two members in Year 3. Visits to schools were scheduled generally every 4 to 6 weeks. Some schools were visited more often if a school or teacher requested assistance or if the project staff or the school district administrators identified a school or teacher to be high-needs. Assistance included planning, teaching/modeling, co-teaching, working with small groups of students during science instruction, delivering supplies, and providing additional resources for all students and specifically for ELLs (e.g., home-learning suggestions, lab management strategies, and supplemental assessment materials). The project staff also met regularly with school administrators during school visits to address any of their concerns.

**Measures, Data Collection, and Data Analysis**

**Measures**

At the beginning of their participation in the project, all teachers in the treatment and control groups completed a project-developed questionnaire that included a section about their views of HSTA (See Appendix B). Four items formed a scale which reported on teachers’ perceptions of the impact of the state’s science test on teaching practices (i.e., ITP). For the items comprising the scale, the ratings included 1 = greatly decreases, 2 = decreases, 3 = increases, and 4 = greatly increases. Three items were analyzed as individual items: (a) the state’s standards in science provide good guidelines (i.e., SGG), (b) the state’s accountability plan provides a reasonable plan for grading schools (i.e., APR), and (c) the state’s science test scores do not accurately reflect ESOL students’ science knowledge (i.e., ESK). For the individual items, the ratings included 1 = strongly
disagree, 2 = somewhat disagree, 3 = somewhat agree, and 4 = strongly agree. See Table 3 for details on the items which comprised the scale.

Table 3

Designation of Outcome Variables

<table>
<thead>
<tr>
<th>Scale / Item</th>
<th>Abbreviation</th>
<th>Reverse Coded</th>
</tr>
</thead>
<tbody>
<tr>
<td>Scale: Impact on Teaching Practices</td>
<td>ITP</td>
<td></td>
</tr>
<tr>
<td>i. Others telling me what to teach in science</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>ii. Others telling me how to teach science</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>iii. Time spent in preparation for testing</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>iv. Drill and practice for testing</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>Item 1: Science standards provide good guidelines</td>
<td>SGG</td>
<td></td>
</tr>
<tr>
<td>Item 2: Accountability policy is reasonable</td>
<td>APR</td>
<td></td>
</tr>
<tr>
<td>Item 3: State test scores do not reflect ESOL students’ science knowledge</td>
<td>ESK</td>
<td></td>
</tr>
</tbody>
</table>

All teachers also completed a background information form (See Appendix C). This form asked for number of years of teaching (YEARS) and number of science courses taken (COURSES), which were used as predictors in the models. COURSES was computed by adding together teacher responses for the number of courses taken in the following topics: methods of teaching science in elementary school, methods of teaching science in secondary school, physical science (physics, chemistry), earth/space science (astronomy, geology), life science (biology, ecology, environmental science, marine science), and other.

School percentages of students who self-identified as Black (BLK), students receiving free/reduced price lunch (FRL), and students who participated in English as second or other language (ESOL) programs were retrieved from the state’s Department of Education, Education Information & Accountability Services website. This data was collected for treatment and control schools for each year of the intervention. BLK, FRL, and ESOL were used as predictors in the models. The intercorrelations among teacher
professional and school contextual variables demonstrated that while the predictor variables were significantly correlated with each other, none were highly correlated (see Table 4).

Table 4

<table>
<thead>
<tr>
<th></th>
<th>YEARS</th>
<th>COURSES</th>
<th>BLK</th>
<th>FRL</th>
<th>ESOL</th>
</tr>
</thead>
<tbody>
<tr>
<td>YEARS</td>
<td>--</td>
<td>.397**</td>
<td>-.006</td>
<td>-.078*</td>
<td>.014</td>
</tr>
<tr>
<td>COURSES</td>
<td>.397**</td>
<td>--</td>
<td>.011</td>
<td>-.108**</td>
<td>-.035</td>
</tr>
<tr>
<td>BLK</td>
<td>-.006</td>
<td>.011</td>
<td>--</td>
<td>.279**</td>
<td>-.602**</td>
</tr>
<tr>
<td>FRL</td>
<td>-.078*</td>
<td>-.108**</td>
<td>.279**</td>
<td>--</td>
<td>.397**</td>
</tr>
<tr>
<td>ESOL</td>
<td>.014</td>
<td>-.035</td>
<td>-.602**</td>
<td>--</td>
<td>--</td>
</tr>
</tbody>
</table>

**Correlation is significant at the 0.01 level (2-tailed).**

*Correlation is significant at the 0.05 level (2-tailed).

**Data Collection**

The proposal to undertake this intervention was submitted for review to the Institutional Review Board (IRB) at the university. On June 11, 2009, the Social and Behavioral Sciences full IRB determined that this study possessed minimal risk to the human participants and the study’s protocol qualified for exemption from IRB review. Data collection was undertaken at a large, urban school district and also required approval by that institution’s review board, the Research Review Committee (RRC). On July 16, 2010, the RRC granted its approval to conduct the research. Prior to beginning data collection, all participants had the confidentiality of their responses explained to them and any misgivings were addressed. Consent forms were collected.

Teachers from the treatment schools completed the two measures at the beginning of the first workshop and during the last workshop of each year of the intervention. Data collection was conducted by the trained staff facilitating the workshops in a controlled environment without a time limit. Teachers from control schools completed the measures at their school sites during the first and last quarter of each school year of the intervention.
under guidance of the trained staff. Control group teachers were given a small stipend as compensation.

When a teacher completed the questionnaire prior to beginning the intervention, the time was coded as T0 (baseline). When a teacher completed the questionnaire at the end of his/her respective first year of participation in the intervention, time was coded as Time 1. Similarly, at the end of his/her respective second year of participation in the intervention, time was coded as Time 2; and at the end of his/her respective third year of participation in the intervention, time was coded as Time 3. Table 5 displays the number of teachers who completed the questionnaire at each time.

Table 5
Patterns of Teacher Questionnaire Data Collection (N = 359)

<table>
<thead>
<tr>
<th>Treatment</th>
<th>Control</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>n</td>
</tr>
<tr>
<td>Group 1</td>
<td>51</td>
</tr>
<tr>
<td>Group 2</td>
<td>32</td>
</tr>
<tr>
<td>Group 3</td>
<td>49</td>
</tr>
<tr>
<td>Group 4</td>
<td>10</td>
</tr>
<tr>
<td>Group 5</td>
<td>0</td>
</tr>
<tr>
<td>Group 6</td>
<td>14</td>
</tr>
<tr>
<td>Group 7</td>
<td>26</td>
</tr>
<tr>
<td>Group 8</td>
<td>1</td>
</tr>
<tr>
<td>Totals</td>
<td>183</td>
</tr>
</tbody>
</table>

A teacher could start participation during any time of the three-year intervention. For example, if a teacher started teaching at a school during Year 3 of the intervention, when the teacher completed the questionnaire at the end of the year, the time would be coded as Time 1 since it was his/her first year of participation in the intervention. The maximum number of time points that a teacher could have is four (Group 1), in which case the teacher completed a baseline questionnaire, then participated in three years of the intervention and completed the questionnaire at the end of each year.
Data Analysis

Data storage and calculation of descriptive statistics were conducted using SPSS 19. Scale scores were computed using the average of the responses to the items that comprised the scale after reverse coding each item. Respondents were excluded if they failed to answer at least three of the four items on the ITP scale. The reliability of the obtained scale scores were estimated using Cronbach’s alpha ($\alpha$). Nunnally and Bernstein (1994) noted that an alpha of 0.70 and above is reliable, while some authors suggested that values starting at 0.67 are acceptable (Cohen, Marion, & Morrison, 2008). Whiston (2005) suggested that Cronbach alphas are usually low and conservative estimates of reliability. The reliability estimates for the scores on the ITP scale for each group (i.e., treatment and control) and at each time point were greater than .80 (see Table 6), which is considered acceptable.

Table 6 displays the means and standard deviations for the ITP scale for both groups at each time point. Table 6 also displays the percent responding at each level for the SGG, APR, and ESK items. The descriptive statistics are briefly discussed below and though it provides some information regarding trends, there is a considerable amount of missing data across time points. As a result, multilevel model (MLM) analyses examining change over time were conducted. The MLM uses full information likelihood (FIML) estimates to handle missing data to produce unbiased estimates.

Table 6
Table 6
Descriptive Statistics for Teachers’ Responses to the Questionnaire
a. Scale Statistics

<table>
<thead>
<tr>
<th></th>
<th>Baseline</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>M</td>
<td>n</td>
<td>SD</td>
<td>M</td>
<td>n</td>
<td>SD</td>
<td>M</td>
<td>n</td>
<td>SD</td>
</tr>
<tr>
<td>Treatment</td>
<td>1.77</td>
<td>121</td>
<td>0.52</td>
<td>1.74</td>
<td>166</td>
<td>0.56</td>
<td>1.81</td>
<td>92</td>
<td>0.62</td>
</tr>
<tr>
<td>Control</td>
<td>1.80</td>
<td>118</td>
<td>0.52</td>
<td>1.81</td>
<td>163</td>
<td>0.59</td>
<td>1.74</td>
<td>85</td>
<td>0.55</td>
</tr>
</tbody>
</table>

b. Item Percentages

<table>
<thead>
<tr>
<th>Rating</th>
<th>Baseline</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td>SGG</td>
<td>Treatment</td>
<td>2.5</td>
<td>9.2</td>
<td>68.3</td>
<td>20.0</td>
<td>2.4</td>
<td>8.3</td>
<td>68.0</td>
<td>21.3</td>
</tr>
<tr>
<td></td>
<td>Control</td>
<td>1.6</td>
<td>9.7</td>
<td>73.4</td>
<td>15.3</td>
<td>1.2</td>
<td>8.7</td>
<td>69.6</td>
<td>20.5</td>
</tr>
<tr>
<td>APR</td>
<td>Treatment</td>
<td>17.9</td>
<td>32.1</td>
<td>47.0</td>
<td>3.0</td>
<td>25.3</td>
<td>31.3</td>
<td>4.2</td>
<td>47.3</td>
</tr>
<tr>
<td></td>
<td>Control</td>
<td>18.2</td>
<td>36.4</td>
<td>35.7</td>
<td>9.8</td>
<td>26.9</td>
<td>35.6</td>
<td>31.3</td>
<td>6.3</td>
</tr>
<tr>
<td>ESK</td>
<td>Treatment</td>
<td>2.6</td>
<td>14.8</td>
<td>47.8</td>
<td>34.8</td>
<td>4.4</td>
<td>9.5</td>
<td>45.6</td>
<td>40.5</td>
</tr>
<tr>
<td></td>
<td>Control</td>
<td>5.0</td>
<td>7.5</td>
<td>49.2</td>
<td>38.3</td>
<td>3.9</td>
<td>10.4</td>
<td>40.9</td>
<td>44.8</td>
</tr>
</tbody>
</table>

*Note.* All responses were based on a 4-point rating system. For the scale, higher scores indicate a more positive perception. The ratings for the items are: 1 = strongly disagree, 2 = somewhat disagree, 3 = somewhat agree, and 4 = strongly agree. Percentages included are valid percentages.

* Items for the scale were reverse coded and the means were calculated with the reverse coded responses.

*Cronbach α = .84 at baseline, T1, T2, and T3,*

*Cronbach α = .84 at baseline, .88 at T1, .85 at T2, and .89 at T3.*
**Descriptives.** At T0, the mean ITP for the treatment ($M = 1.77, SD = 0.52$) and control ($M = 1.80, SD = 0.52$) group were similar. The mean ITP for the treatment group increased slightly from T1 ($M = 1.74, SD = 0.56$) to T2 ($M = 1.81, SD = 0.62$) and T3 ($M = 1.96, SD = 0.55$). The mean ITP for the control group decreased slightly from T1 ($M = 1.81, SD = 0.59$) to T2 ($M = 1.74, SD = 0.55$) and T3 ($M = 1.78, SD = 0.65$).

At T0, agreement with SGG was fairly similar for the treatment (68% somewhat agreed and 20% strongly agreed) and control groups (73% somewhat agreed and 15% strongly agreed). Across time, the treatment group increased, particularly in the somewhat agree category from T1 (68% somewhat agreed and 21% strongly agreed) to T2 (70% somewhat agreed and 21% strongly agreed) and T3 (74% somewhat agreed and 20% strongly agreed). In the control group, the percentage of teachers who somewhat agreed fluctuated, while those in strong agreement decreased from T1 (70% somewhat agreed and 21% strongly agreed) to T2 (75% somewhat agreed and 13% strongly agreed) and T3 (69% somewhat agreed and 12% strongly agreed).

At T0, agreement with APR was different in the treatment group (47% somewhat agreed and 3% strongly agreed) and control group (36% somewhat agreed and 10% strongly agreed). Across time, the treatment group decreased in agreement from T1 (31% somewhat agreed and 4% strongly agreed) to T2 (26% somewhat agreed and 1% strongly agreed) and T3 (30% somewhat agreed and 2% strongly agreed). In the control group, the percentage of teachers who somewhat agreed also decreased from T1 (31% somewhat agreed and 6% strongly agreed) to T2 (25% somewhat agreed and 4% strongly agreed) and T3 (24% somewhat agreed and 2% strongly agreed).
At T0, agreement with ESK was fairly similar for the treatment (48% somewhat agreed and 35% strongly agreed) and control groups (50% somewhat agreed and 38% strongly agreed). Across time, the treatment group increased in agreement from T1 (46% somewhat agreed and 41% strongly agreed) to T2 (40% somewhat agreed and 47% strongly agreed) and T3 (57% somewhat agreed and 39% strongly agreed). In the control group, the percentage of teachers in agreement also increased from T1 (41% somewhat agreed and 45% strongly agreed) to T2 (47% somewhat agreed and 43% strongly agreed) and T3 (43% somewhat agreed and 50% strongly agreed).

As a first step, the pattern of change on the measures was examined. Three patterns of change were considered: a) linear change, with a constant rate of change across all time points, b) piecewise change with different rates of change occurring from T0 to T1 and T1 to T3, and c) piecewise change with different rates of change occurring from T0 to T2 and T2 to T3. Separate models were examined for the treatment and control groups specifying the three different patterns of change and compared by calculating both Akaike information criteria (AIC) and Bayesian information criteria (BIC). Both the AIC and BIC allow competing models to be compared for fit; with smaller values indicating a better fit. AIC and BIC values did not differ greatly and indicated that the piecewise model with different rates of change occurring from T0 to T2 and T2 to T3 fit both the treatment and control groups across outcome variables. Since the growth patterns for the treatment and control teachers were similar, one model was built for each outcome variable with GROUP as a predictor variable.

**MLM analysis.** Although empirical research has focused on answering questions about change, it is only since the 1980s that statistical models capable of studying change
were developed (Singer & Willett, 2003). This class of statistical models uses longitudinal data and is known by different names, including individual growth models, multilevel models, and hierarchical linear models. In this study it is referred to as MLM and was selected primarily because of its value to examine change over time.

One benefit of MLM is that it estimates both initial status on the variable of interest and rate of change on the variable of interest free from measurement error variance. Other methods analyze only mean change, while ignoring differences in initial status and treating differences in change as error variance (Raudenbush & Bryk, 2002).

Another benefit of using MLM is that it allows unbalanced data sets. In this study, there are missing data not only from respondents who did not answer all items on the measures, but also due to teacher attrition that could impact results. Possible reasons for attrition included administrative reassignment, transfers from one school/district/state to another, and leaving the teaching profession permanently. If attrition is random, individual participants are not required to have the same number of data collection points. By using MLM, all participants even those with only one data collection time point are included in the analysis. To account for missing data points, MLM uses full information maximum likelihood estimation to compute parameter estimates based on the data collected for each participant (Singer & Willett, 2003). The MLM analyses were conducted using HLM 7.0 software to examine how the outcome variables changed over time and identify variations in the outcome variables by teacher professional variables and school contextual factors.

**ITP scale.** The model building process imposed to examine the ITP scale included first estimating a two-level unconditional model with time nested within
teachers to determine the intraclass correlation coefficient (ICC). Consistent with Raudenbush and Bryk (2002), ICC was computed using the following formula: 

$$ICC = \frac{\tau_{00}}{\tau_{00} + \sigma^2}$$

where $\tau_{00}$ is the estimated Level-2 variance for the model and $\sigma^2$ is the estimated Level-1 variance for the model.

Next, the Level-1 model was built. Since teachers’ perceptions showed piecewise growth, the models for each outcome variable was specified such that two separate linear growth factors were modeled. Time variables were created for a piecewise model estimating growth from T0 to T2 and T2 to T3 as presented in Table 7.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Range</th>
<th>T0</th>
<th>T1</th>
<th>T2</th>
<th>T3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Y1</td>
<td>T0 to T2</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>Y2</td>
<td>T2 to T3</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

Time variables Y1 and Y2 were entered as uncentered predictors.

Then, the Level-2 models were developed. Six predictor variables were entered in the model at Level-2 as predictors of the intercept (i.e., baseline). The predictors included GROUP which was entered uncentered, since it is a dichotomous variable; two teacher professional variables (1) years of teaching (YEARS) and (2) number of science courses taken (COURSES), which were entered grand mean centered such that the intercept would reflect the average across all teachers; and three school contextual factors (1) percent of Black students (BLK), (2) percent of low SES students (FRL), and (3) percent of limited English proficient students (ESOL), which were entered grand mean centered such that the intercept would reflect the average across all schools. Percent of ESE was not included as a predictor because this school contextual factor demonstrated minimal variation across participating schools. GROUP, YEARS, COURSES, BLK, FRL, and
ESOL were not included in the model as predictors of the slope of primary or secondary change since there was not a significant amount of variation between individuals in either slope (i.e., the variance of the parameter was not statistically significant). The following was the full model used in the analysis of the ITP scale.

Level-1 Model

\[ Y_{ti} = \pi_{0i} + \pi_{1i}(Y_{1ti}) + \pi_{2i}(Y_{2ti}) + e_{ti} \]

As illustrated in the equation above, \( \pi_{0i} \) represents teacher \( i \)'s true value of ITP at baseline, \( \pi_{1i} \) represents teacher \( i \)'s rate of change in true value of ITP from baseline to year 2, and \( \pi_{2i} \) represents teacher \( i \)'s rate of change in true value of ITP from year 2 to year 3.

Level-2 Model

\[ \pi_{0i} = \beta_{00} + \beta_{01}(GROUP_i) + \beta_{02}(YEARS_i) + \beta_{03}(COURSES_i) + \beta_{04}(BLK_i) + \beta_{05}(FRL_i) + \beta_{06}(ESOL_i) + r_{0i} \]

\[ \pi_{1i} = \beta_{10} + r_{1i} \]

\[ \pi_{2i} = \beta_{20} + r_{2i} \]

As illustrated in the equation above, \( \beta_{00} \) represents the average across all teachers of true ITP at baseline for control teachers, after controlling for GROUP, YEARS, COURSES, BLK, FRL, and ESOL. \( \beta_{01} \) represents the difference in average across all teachers of true ITP at baseline between treatment and control teachers, when other variables are controlled. \( \beta_{02} \) represents the difference in average across all teachers of true ITP at baseline for each additional year of teaching, when other variables are controlled. \( \beta_{03} \) represents the difference in average across all teachers of true ITP at baseline for each additional science course taken, when other variables are controlled. \( \beta_{04} \) represents the difference in average across all teachers of true ITP at baseline for each additional percentage of students who self-identified as Black at a teacher’s school, when other
variables are controlled. \( \beta_{05} \) represents the difference in average across all teachers of true ITP at baseline for each additional percentage of students who received free/reduced lunch at a teacher’s school, when other variables are controlled. \( \beta_{06} \) represents the difference in average across all teachers of true ITP at baseline for each additional percentage of students who participated in ESOL programs at a teacher’s school, when other variables are controlled. \( \beta_{10} \) represents the average annual rate of true change in ITP from baseline to year 2 for control teachers. \( B_{20} \) represents the average annual rate of true change in ITP from year 2 to year 3 for control teachers.

**SGG, APR, and ESK items.** A continuous outcome model may not be capable of measuring enough variation recorded by a rating scale with ratings 1 = strongly disagree, 2 = somewhat disagree, 3 = somewhat agree, and 4 = strongly agree such as that used to measure the three individual items in this study (i.e., SGG, APR, and ESK). Rating scales result in a non-normal outcome variable and continuous outcome models assume random effects are normally distributed (Raudenbush and Bryk, 2002). O’Connell and McCoach (2008) and Raudenbush and Bryk (2002) suggested treating such research data not only as obtained through a natural hierarchy, but when possible as ordinal scale variables which result in normal distributions. A multilevel model for ordinal response data was used to analyze the three individual items SGG, APR, and ESK.

The data had four possible ordered categories, based on four response ratings 1 = strongly disagree, 2 = somewhat disagree, 3 = somewhat agree, and 4 = strongly agree. The cumulative probabilities of a response of 1 versus 2, 3, or 4; of a response of 1 or 2 versus 3 or 4; and a response of 1, 2, or 3 versus 4 were calculated. The probability of a response of 1, 2, 3, or 4 is known to be 1.0 and was not included. Cumulative
probabilities were used to compute cumulative odds which represent the odds that a response would most likely be at or below a threshold or cut point rather than beyond it. The natural log of the cumulative odds, the logit, is a quotient that compares the probability of a response to the probability of another response and was also calculated. Since the data were partitioned into three sequential categories of interest, three thresholds or cut points resulted in the calculation of the following comparisons: (1) the probability of the $i^{th}$ teacher responding 1 compared to responding 2, 3, or 4; (2) the probability of the $i^{th}$ teacher responding 1 or 2 compared to 3 or 4; and (3) the probability of the $i^{th}$ teacher responding 1, 2, or 3 compared to 4.

The model building for the three individual items followed similar development than that for the scale model. First, a two-level unconditional model with time nested within teachers was estimated to determine ICC. The ICC was calculated using the formula $\text{ICC} = \frac{\tau_{00}}{\tau_{00} + 3.29}$ where 3.29 corresponds to $\frac{\pi^2}{3}$ because this variance represents the within-group variance for ICC calculations for dichotomous data and the ICC can be similarly defined for ordinal outcomes (Snijders & Bosker, 1999).

Next, the Level-1 model was developed. Proportional odds analysis was used with four categories and three thresholds or cut points which examined the following comparisons: (i) the probability of the $i^{th}$ teacher strongly disagreeing compared to somewhat disagreeing, somewhat agreeing, or strongly agreeing; (ii) the probability of the $i^{th}$ teacher strongly disagreeing or somewhat disagreeing compared to somewhat agreeing or strongly agreeing; and (iii) the probability of the $i^{th}$ teacher strongly disagreeing, somewhat disagreeing, or somewhat agreeing compared to strongly agreeing.
Then, the Level-2 model was developed. The same six predictor variables
described for the scale model – GROUP, YEARS, COURSES, BLK, FRL, and ESOL –
were entered at Level-2 as predictors of the intercept (i.e., baseline). GROUP was entered
uncentered and YEARS, COURSES, BLK, FRL, and ESOL were entered centered
around the grand mean. GROUP, YEARS, COURSES, BLK, FRL, and ESOL were not
included in the model as predictors of the slope of primary or secondary change since
there was not a significant amount of variation between individuals in either slope (i.e.,
the variance of the parameter was not statistically significant). The following was the full
model used in the analysis of each of the three individual items. TOV represents the
outcome variable being tested (i.e., SGG, APR, or ESK).

Level-1 Model

\[
\begin{align*}
\text{Prob}[R_{i} <= 1|\pi_i] &= \phi_{1i} = \phi_{1i} \\
\text{Prob}[R_{i} <= 2|\pi_i] &= \phi_{2i} = \phi_{1i} + \phi_{2i} \\
\text{Prob}[R_{i} <= 3|\pi_i] &= \phi_{3i} = \phi_{1i} + \phi_{2i} + \phi_{3i} \\
\text{Prob}[R_{i} <= 4|\pi_i] &= 1.0 \\
\phi_{1i} &= \text{Prob}[TOV(1) = 1|\pi_i] \\
\phi_{2i} &= \text{Prob}[TOV(2) = 1|\pi_i] \\
\phi_{3i} &= \text{Prob}[TOV(3) = 1|\pi_i]
\end{align*}
\]

In the equations below, $\phi_{1i}$ represents the cumulative odds that the $i^{th}$ teacher will
strongly disagree with TOV, compared to somewhat disagree, somewhat agree, or
strongly agree. $\phi_{2i}$ represents the cumulative odds that the $i^{th}$ teacher will strongly
disagree or somewhat disagree with TOV, compared to somewhat agree or strongly
agree; and $\phi_{3i}$ represents the cumulative odds that the $i^{th}$ teacher will strongly disagree,
somewhat disagree, or somewhat agree with TOV, compared to strongly agree.

\[
\begin{align*}
\log[\phi_{1i}/(1 - \phi_{1i})] &= \pi_{0i} + \pi_{1i}*(Y_{1i}) + \pi_{2i}*(Y_{2i}) \\
\log[\phi_{2i}/(1 - \phi_{2i})] &= \pi_{0i} + \pi_{1i}*(Y_{1i}) + \pi_{2i}*(Y_{2i}) + \delta_2 \\
\log[\phi_{3i}/(1 - \phi_{3i})] &= \pi_{0i} + \pi_{1i}*(Y_{1i}) + \pi_{2i}*(Y_{2i}) + \delta_3
\end{align*}
\]
Level-2 Model

\[ \pi_{0i} = \beta_{00} + \beta_{01}(\text{GROUP}) + \beta_{02}(\text{YEARS}) + \beta_{03}(\text{COURSES}) + \beta_{04}(\text{BLK}) + \beta_{05}(\text{FRL}) + \beta_{06}(\text{ESOL}) + r_{0i} \]

\[ \pi_{1i} = \beta_{10} + r_{1i} \]

\[ \pi_{2i} = \beta_{20} + r_{2i} \delta_2 + \delta_3 \]

As illustrated in the equation above, \( \beta_{00} \) represents the average across all teachers of true TOV at baseline for control teachers, after controlling for GROUP, YEARS, COURSES, BLK, FRL, and ESOL. \( \beta_{01} \) represents the difference in average across all teachers of true TOV at baseline between treatment and control teachers, when other variables are controlled. \( \beta_{02} \) represents the difference in average across all teachers of true TOV at baseline for each additional year of teaching, when other variables are controlled. \( \beta_{03} \) represents the difference in average across all teachers of true TOV at baseline for each additional science course taken, when other variables are controlled. \( \beta_{04} \) represents the difference in average across all teachers of true TOV at baseline for each additional percentage of students who self-identified as Black at a teacher’s school, when other variables are controlled. \( \beta_{05} \) represents the difference in average across all teachers of true TOV at baseline for each additional percentage of students who received free/reduced lunch at a teacher’s school, when other variables are controlled. \( \beta_{06} \) represents the difference in average across all teachers of true TOV at baseline for each additional percentage of students who participated in ESOL programs at a teacher’s school, when other variables are controlled. \( \beta_{10} \) represents the average annual rate of true change in TOV from baseline to year 2 for control teachers. \( \beta_{20} \) represents the average annual rate of true change in TOV from year 2 to year 3 for control teachers.
CHAPTER 5

Results

The purpose of this research was to examine elementary science teachers’ perceptions of HSTA prior to an intervention (research question 1), and if participation in the intervention changed the teachers’ perceptions of HSTA (research question 3a and 3b). In addition, teacher professional variables and school contextual factors were examined as predictors of initial perceptions of HSTA (research question 2) and change in perceptions over time (research questions 4a and 4b). In this section, results for each of the research questions are organized with regard to the ITP scale followed by each of the three individual items.

Impact on Teaching Practices (ITP)

Unconditional model. Results for the ITP scale are displayed in Table 8. The Level-1 variance ($\sigma^2$) for the unconditional model was 0.25, and the Level-2 variance ($\tau_{00}$) was 0.14. The ICC is .37 which indicates that 37% of the variance in ITP was due to interindividual teacher differences (Level-2), while 63% of the variance in ITP was due to intraindividual differences (Level-1). This suggests that there is more variation over time than variation between teachers. The variance of the intercept ($r_0$) was statistically significant ($\chi^2[318] = 762.68, p < .001$), suggesting a statistically significant amount of variance in ITP scores between teachers. The Level-1 coefficient for initial ITP status was 1.90 ($t[312] = 45.33, p < .001$), indicating that across all teachers, the average starting ITP score was 1.90 (research question 1).
Table 8
Results for ITP Scale

<table>
<thead>
<tr>
<th>Fixed Effect</th>
<th>Unconditional Model</th>
<th></th>
<th></th>
<th></th>
<th>Change Model</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th>Full Model</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Coeff.</td>
<td>SE</td>
<td>t-ratio</td>
<td>d.f.</td>
<td>p-value</td>
<td>Coeff.</td>
<td>SE</td>
<td>t-ratio</td>
<td>d.f.</td>
<td>p-value</td>
<td>Coeff.</td>
<td>SE</td>
<td>t-ratio</td>
</tr>
<tr>
<td>For INTBPT1, πc</td>
<td>1.89</td>
<td>0.03</td>
<td>65.79</td>
<td>318</td>
<td>&lt;0.001</td>
<td>1.88</td>
<td>0.03</td>
<td>62.82</td>
<td>318</td>
<td>&lt;0.001</td>
<td>1.90</td>
<td>0.04</td>
<td>45.33</td>
</tr>
<tr>
<td>INTBPT2, β00</td>
<td>-0.03</td>
<td>0.06</td>
<td>-0.48</td>
<td>312</td>
<td>0.632</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GROUP, β01</td>
<td>0.00</td>
<td>0.00</td>
<td>-0.26</td>
<td>312</td>
<td>0.799</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>YEARS, β02</td>
<td>0.00</td>
<td>0.00</td>
<td>1.12</td>
<td>312</td>
<td>0.262</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>COURSES, β03</td>
<td>0.00</td>
<td>0.00</td>
<td>0.47</td>
<td>312</td>
<td>0.640</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BLK, β04</td>
<td>0.00</td>
<td>0.00</td>
<td>0.03</td>
<td>312</td>
<td>0.979</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FRL, β05</td>
<td>0.00</td>
<td>0.00</td>
<td>0.54</td>
<td>312</td>
<td>0.592</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ESOL, β06</td>
<td>0.00</td>
<td>0.00</td>
<td>0.96</td>
<td>318</td>
<td>0.337</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>For Y1 slope, π1</td>
<td>0.07</td>
<td>0.07</td>
<td>0.94</td>
<td>318</td>
<td>0.35</td>
<td>0.07</td>
<td>0.07</td>
<td>0.96</td>
<td>318</td>
<td>0.337</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>INTBPT2, β10</td>
<td>0.01</td>
<td>0.05</td>
<td>0.17</td>
<td>318</td>
<td>0.86</td>
<td>0.01</td>
<td>0.05</td>
<td>0.23</td>
<td>318</td>
<td>0.818</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Estimates of Variance Components</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random Effect</td>
</tr>
<tr>
<td>INTBPT1, r0</td>
</tr>
<tr>
<td>Y1 slope, r1</td>
</tr>
<tr>
<td>Y2 slope, r2</td>
</tr>
<tr>
<td>level-1, e</td>
</tr>
</tbody>
</table>
Change model. The Level-1 coefficient for GROUP was -0.03 and was not statistically significant ($t_{[312]} = -0.48, p = .632$). This indicates that ITP scores did not vary significantly across treatment and control groups. The coefficient for primary change in ITP was 0.07 ($t_{[318]} = 0.96, p = .337$), indicating that across all teachers, ITP increased by 0.07 points, on average, from T0 to T2. The coefficient for secondary change in ITP was 0.01 ($t_{[318]} = 0.23, p = .818$), indicating that across all teachers, ITP increased by 0.01 points, on average, from T2 to T3. The estimate of the variance component for the Y1 intercept was 0.00, which was not significant, suggesting that there was not a significant amount of between teacher variation around the initial true change in ITP from T0 to T2. Also, the estimate of the variance component for the Y2 intercept was 0.01 indicating no significant amount of between teacher variation around the initial true change in ITP from T2 to T3. The results suggest that there was no significant difference in ITP for the treatment and control groups and that if change took place from T0 to T2 or from T2 to T3 all teachers changed in the same way over time (research question 3a and 3b). Because teachers’ perceptions did not differ over time (i.e., the variance of Y1 and Y2 was not significant), predictor variables were not added to the model as predictors of the slope of primary or secondary change for ITP (research question 4a and 4b).

Full model. Six predictor variables, GROUP, YEARS, COURSES, BLK, FRL, and ESOL were included in this model as predictors of the intercept. Fixed effects show that none of the predictors were statistically significant; suggesting that none of the variables tested predicted ITP at baseline (research question 2). Residual variance
remains, however, in the intercept for ITP across teachers. Other predictors could be considered in future models.

**Science Standards Provide Good Guidelines (SGG)**

**Unconditional model.** Results for the SGG ordinal model analysis are presented in Table 9. The ICC coefficient is .41, which suggests that 41% of the variability in SGG is due to between teacher differences and 59% is attributable to within teacher differences. With no explanatory variables in the model, the cumulative logit prediction of SGG response at or below strongly disagree on average across teachers at baseline is -4.58 and steadily increases across the cut points for SGG response at or below strongly agree by 1.86 (-4.58 + 6.44 = 1.86) for comparison category teachers strongly disagree versus teachers somewhat disagree, somewhat agree, or strongly agree. The cumulative logit prediction of SGG response at or below somewhat disagree on average across teachers at baseline is 1.99 and steadily increases across the cut points for SGG response below strongly agree by 8.43 (1.99 + 6.44 = 1.86) for comparison category teachers strongly disagree or somewhat disagree versus teachers somewhat agree or strongly agree. The cumulative logit prediction of SGG response below somewhat agree on average across teachers at baseline is 6.44 and increases across the cut points for SGG response below strongly agree by 12.88 (6.44 + 6.44 = 12.88) for comparison category strongly disagree, somewhat disagree, or somewhat agree versus strongly agree. Transforming the predicted cumulative logit to a proportional odds and then to cumulative probabilities results in $P(R_{ij} \leq 1) = .01$. The logits estimated from this model show that the variance between teachers ($\tau_{00} = 2.32$, $p < .01$) is statistically significant
and suggests that across all teachers, the most likely baseline SGG response was 2.32 (research question 1).

**Change model.** This model included two time parameters (Y1 and Y2). Fixed effects and variance components were not statistically significant for either parameter. These results suggest that if between teacher variation around the initial true change in SGG took place from T0 to T2 or from T2 to T3 all teachers changed in the same way over time (research question 3a and 3b). As a result, predictor variables were not included in either of the slope estimates (research question 4a and 4b).

**Full model.** Six predictor variables, GROUP, YEARS, COURSES, BLK, FRL, and ESOL were included in this model as predictors of the intercept. GROUP shows that there was no difference in APR at baseline for treatment and control teachers. Fixed effects show that COURSES, BLK, and FRL were not statistically significant predictors of SGG at baseline for treatment and control teachers and that YEARS is a statistically significant predictor ($\text{Coeff} = -0.05; \text{SE} = .01; \text{OR} = 0.95$). This suggests that for each additional year of teaching, the odds that a teacher at baseline will strongly agree that standards provided good guidelines versus the combined somewhat agree, somewhat disagree and strongly disagree are 5% less likely given the other variables are held constant in the model; likewise for each additional year of teaching, the odds that a teacher at baseline will strongly agree or somewhat agree versus the combined somewhat disagree and strongly disagree are 5% less likely given the other variables are held constant in the model; and for each additional year of teaching experience, the odds that a teacher at baseline will strongly agree, somewhat agree or somewhat disagree versus strongly disagree are 5% less likely given the other variables are held constant in the
model. As the number of years teaching increases, the likelihood of being at or above a cut point increases (research question 2). Residual variance remains, however, in the intercept for SGG across teachers. Other predictors could be considered in future modeling.

Table 9

Results for SGG Ordinal Model (Proportional Odds)

<table>
<thead>
<tr>
<th>Fixed Effect</th>
<th>Unconditional Model</th>
<th>Change Model</th>
<th>Full Model</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Coeff. (SE)</td>
<td>OR</td>
<td>Coeff. (SE)</td>
</tr>
<tr>
<td>For INTRCPT1 slope, $\pi_0$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>INTRCPT2, $\beta_{00}$</td>
<td>-4.58 (.29)</td>
<td>0.01**</td>
<td>-4.63 (0.30)</td>
</tr>
<tr>
<td>GROUP, $\beta_{01}$</td>
<td>-0.21 (.25)</td>
<td>0.81</td>
<td>-0.21 (.25)</td>
</tr>
<tr>
<td>YEARS, $\beta_{02}$</td>
<td>-0.05 (.01)</td>
<td>0.95**</td>
<td>-0.05 (.01)</td>
</tr>
<tr>
<td>COURSES, $\beta_{03}$</td>
<td>0.02 (.02)</td>
<td>1.03</td>
<td>0.02 (.02)</td>
</tr>
<tr>
<td>BLK, $\beta_{04}$</td>
<td>-0.01 (.01)</td>
<td>0.99</td>
<td>-0.01 (.01)</td>
</tr>
<tr>
<td>FRL, $\beta_{05}$</td>
<td>0.00 (.01)</td>
<td>1.00</td>
<td>0.00 (.01)</td>
</tr>
<tr>
<td>ESOL, $\beta_{06}$</td>
<td>-0.01 (.01)</td>
<td>0.99</td>
<td>-0.01 (.01)</td>
</tr>
<tr>
<td>For Y1 slope, $\pi_1$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>INTRCPT2, $\beta_{10}$</td>
<td>0.01 (.33)</td>
<td>1.01</td>
<td>0.03 (.33)</td>
</tr>
<tr>
<td>For Y2 slope, $\pi_2$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>INTRCPT2, $\beta_{20}$</td>
<td>0.16 (.22)</td>
<td>1.17</td>
<td>0.15 (.22)</td>
</tr>
<tr>
<td>For THOLD2, $\delta_2$</td>
<td>1.99 (.25)</td>
<td>7.30**</td>
<td>1.99 (.26)</td>
</tr>
<tr>
<td>For THOLD3, $\delta_3$</td>
<td>6.44 (.30)</td>
<td>625.01**</td>
<td>6.45 (.30)</td>
</tr>
</tbody>
</table>

Random Effects

<table>
<thead>
<tr>
<th></th>
<th>Variance</th>
<th>Variance</th>
<th>Variance</th>
</tr>
</thead>
<tbody>
<tr>
<td>INTRCPT1, $r_0$</td>
<td>2.32**</td>
<td>2.50**</td>
<td>2.26**</td>
</tr>
<tr>
<td>Y1 slope, $r_1$</td>
<td>0.04</td>
<td>0.04</td>
<td></td>
</tr>
<tr>
<td>Y2 slope, $r_2$</td>
<td>0.06</td>
<td>0.05</td>
<td></td>
</tr>
</tbody>
</table>

Notes: FPQL estimation; group-mean centering of YEARS, COURSES, BLK, FRL, and ESOL; **p<.01

Accountability Policy is Reasonable (APR)

Unconditional model. Results for the APR ordinal model analysis are presented in Table 10. The ICC is .31 which suggests that 31% of the variability in APR is due to between teacher differences and 69% is attributable to within teacher differences. With no explanatory variables in the model, the cumulative logit prediction of APR response at
or below strongly disagree on average across teachers at baseline is -1.06 and steadily increases across the cut points for APR response at or below strongly agree by 3.64 (-1.06 + 4.70 = 3.64) for comparison category teachers strongly disagree versus teachers somewhat disagree, somewhat agree, or strongly agree. The cumulative logit prediction of APR response at or below somewhat disagree on average across teachers at baseline is 1.69 and steadily increases across the cut points for APR response below strongly agree by 6.39 (1.69 + 4.70 = 6.39) for comparison category teachers strongly disagree or somewhat disagree versus teachers somewhat agree or strongly agree. The cumulative logit prediction of APR response below somewhat agree on average across teachers at baseline is 4.70 and increases across the cut points for SGG response below strongly agree by 9.40 (4.70 + 4.70 = 9.40) for comparison category strongly disagree, somewhat disagree, or somewhat agree versus strongly agree. Transforming the predicted cumulative logit to a proportional odds and then to cumulative probabilities results in

\[ P(R_{ij} \leq 1) = .35. \]

The logits estimated from this model show that there is statistically significant variance between teachers (\(\tau_{00} = 1.50, p < .01\)) and suggests that across all teachers, the most likely baseline APR response was 1.50 (research question 1).

**Change model.** This model included two time parameters (Y1 and Y2). Fixed effects and variance components were not statistically significant for either parameter. These results suggest that if between teacher variation around the initial true change in APR took place from T0 to T2 or from T2 to T3 all teachers changed in the same way over time (research question 3a and 3b). As a result, predictor variables were not included in either of the slope estimates (research question 4a and 4b).
**Full model.** Six predictor variables, GROUP, YEARS, COURSES, BLK, FRL, and ESOL were included in this model as predictors of the intercept. GROUP shows that there was no difference in APR at baseline for treatment and control teachers. Fixed effects show that none of the predictors were statistically significant (research question 2). Residual variance remains, however, in the intercept for APR across teachers. Other predictors could be considered in future models.

Table 10

*Results for APR Ordinal Model (Proportional Odds)*

<table>
<thead>
<tr>
<th>Fixed Effect</th>
<th>Unconditional Model</th>
<th>Change Model</th>
<th>Full Model</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Coeff. (SE)</td>
<td>OR</td>
<td>Coeff. (SE)</td>
</tr>
<tr>
<td>For INTRCPT1 slope, $\pi_0$</td>
<td>INTRCPT2, $\beta_{00}$</td>
<td>-1.06 (.29)</td>
<td>0.35**</td>
</tr>
<tr>
<td></td>
<td>GROUP, $\beta_{01}$</td>
<td>0.22 (.21)</td>
<td>1.24</td>
</tr>
<tr>
<td></td>
<td>YEARS, $\beta_{02}$</td>
<td>0.01 (.01)</td>
<td>1.02</td>
</tr>
<tr>
<td></td>
<td>COURSES, $\beta_{03}$</td>
<td>0.01 (.01)</td>
<td>1.01</td>
</tr>
<tr>
<td></td>
<td>BLK, $\beta_{04}$</td>
<td>-0.01 (.01)</td>
<td>0.99</td>
</tr>
<tr>
<td></td>
<td>FRL, $\beta_{05}$</td>
<td>0.02 (.01)</td>
<td>1.02</td>
</tr>
<tr>
<td></td>
<td>ESOL, $\beta_{06}$</td>
<td>-0.01 (.01)</td>
<td>0.99</td>
</tr>
<tr>
<td>For Y1 slope, $\pi_1$</td>
<td>INTRCPT2, $\beta_{10}$</td>
<td>-0.47 (.27)</td>
<td>0.63</td>
</tr>
<tr>
<td>For Y2 slope, $\pi_0$</td>
<td>INTRCPT2, $\beta_{20}$</td>
<td>0.90 (.18)</td>
<td>2.45</td>
</tr>
<tr>
<td>For THOLD2, $\delta_2$</td>
<td>1.69 (.09)</td>
<td>5.41**</td>
<td></td>
</tr>
<tr>
<td>For THOLD3, $\delta_3$</td>
<td>4.70 (.21)</td>
<td>109.65**</td>
<td></td>
</tr>
</tbody>
</table>

Random Effects

<table>
<thead>
<tr>
<th></th>
<th>Variance</th>
<th>Variance</th>
<th>Variance</th>
</tr>
</thead>
<tbody>
<tr>
<td>INTRCPT1, $r_0$</td>
<td>1.50**</td>
<td>1.86**</td>
<td>1.84**</td>
</tr>
<tr>
<td>Y1 slope, $r_1$</td>
<td>0.08</td>
<td>0.10</td>
<td>0.10</td>
</tr>
<tr>
<td>Y2 slope, $r_2$</td>
<td>0.26</td>
<td>0.27</td>
<td>0.27</td>
</tr>
</tbody>
</table>

Notes: FPQL estimation; group-mean centering of YEARS, COURSES, BLK, FRL, and ESOL; **p<.01

**State Test Scores Do Not Reflect ESOL Students’ Science Knowledge (ESK)**

**Unconditional model.** Results for the ESK ordinal model analysis are presented in Table 11. The intraclass correlation coefficient is .30. This suggests that 30% of the
variability in ESK is due to between teacher differences and 70% is attributable to within teacher differences. With no explanatory variables in the model, the cumulative logit prediction of ESK response at or below strongly disagree on average across teachers at baseline is -3.78 and steadily increases across the cut points for ESK response at or below strongly agree by 0.43 (-3.78 + 4.21 = 0.43) for comparison category teachers strongly disagree versus teachers somewhat disagree, somewhat agree, or strongly agree. The cumulative logit prediction of ESK response at or below somewhat disagree on average across teachers at baseline is 1.56 and steadily increases across the cut points for ESK response below strongly agree by 5.77 (1.56 + 4.21 = 5.77) for comparison category teachers strongly disagree or somewhat disagree versus teachers somewhat agree or strongly agree. The cumulative logit prediction of ESK response below somewhat agree on average across teachers at baseline is 4.21 and increases across the cut points for ESK response below strongly agree by 8.42 (4.21 + 4.21 = 8.42) for comparison category strongly disagree, somewhat disagree, or somewhat agree versus strongly agree.

Transforming the predicted cumulative logit to a proportional odds and then to cumulative probabilities results in $P(R_{ij} \leq 1) = .01$. The logits estimated from this model show that the variance between teachers is statistically significant ($\tau_{00} = 1.42, p < .01$) and suggest that across all teachers, the most likely baseline ESK response was 1.42 (research question 1).

**Change model.** This model included two time parameters (Y1 and Y2). Fixed effects and variance components were not statistically significant for either parameter. These results suggest that if between teacher variation around the initial true change in ESK took place from T0 to T2 or from T2 to T3 all teachers changed in the same way
over time (research question 3a and 3b). As a result, predictor variables were not included in either of the slope estimates (research question 4a and 4b).

Table 11
Results for ESK Ordinal Model (Proportional Odds)

<table>
<thead>
<tr>
<th>Fixed Effect</th>
<th>Unconditional Model</th>
<th>Change Model</th>
<th>Full Model</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Coeff. (SE) OR</td>
<td>Coeff. (SE) OR</td>
<td>Coeff. (SE) OR</td>
</tr>
<tr>
<td>For INTRCPT1 slope, $\pi_0$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>INTRCPT2, $\beta_{00}$</td>
<td>-3.78 (.23) 0.02**</td>
<td>-3.72 (0.24) 0.02**</td>
<td>-3.80 (.26) 0.02**</td>
</tr>
<tr>
<td>GROUP, $\beta_{01}$</td>
<td>0.15 (.21) 1.16</td>
<td></td>
<td></td>
</tr>
<tr>
<td>YEARS, $\beta_{02}$</td>
<td>0.02 (.01) 1.02</td>
<td></td>
<td></td>
</tr>
<tr>
<td>COURSES, $\beta_{03}$</td>
<td>-0.02 (.01) 0.98</td>
<td></td>
<td></td>
</tr>
<tr>
<td>BLK, $\beta_{04}$</td>
<td>0.00 (.01) 0.99</td>
<td></td>
<td></td>
</tr>
<tr>
<td>FRL, $\beta_{05}$</td>
<td>0.00 (.01) 1.00</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ESOL, $\beta_{06}$</td>
<td>-0.01 (.01) 0.99</td>
<td></td>
<td></td>
</tr>
<tr>
<td>For Y1 slope, $\pi_1$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>INTRCPT2, $\beta_{10}$</td>
<td>-0.16 (.29) 0.85</td>
<td>-0.19 (.29) 0.82</td>
<td></td>
</tr>
<tr>
<td>For Y2 slope, $\pi_2$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>INTRCPT2, $\beta_{20}$</td>
<td>-0.37 (.19) 0.69</td>
<td>-0.38 (.20) 0.68</td>
<td></td>
</tr>
<tr>
<td>For THOLD2, $\delta_2$</td>
<td>1.56 (.20) 4.78**</td>
<td>1.59 (.20) 7.32**</td>
<td>1.60 (.20) 4.96**</td>
</tr>
<tr>
<td>For THOLD3, $\delta_3$</td>
<td>4.21 (.23) 67.15**</td>
<td>4.28 (.23) 2.08**</td>
<td>4.31 (.23) 74.29**</td>
</tr>
</tbody>
</table>

Random Effects

<table>
<thead>
<tr>
<th></th>
<th>Variance</th>
<th>Variance</th>
<th>Variance</th>
</tr>
</thead>
<tbody>
<tr>
<td>INTRCPT1, $r_0$</td>
<td>1.42**</td>
<td>1.56**</td>
<td>1.39**</td>
</tr>
<tr>
<td>Y1 slope, $r_1$</td>
<td>0.04</td>
<td>0.02</td>
<td></td>
</tr>
<tr>
<td>Y2 slope, $r_2$</td>
<td>0.10</td>
<td>0.18</td>
<td></td>
</tr>
</tbody>
</table>

Notes: FPQL estimation; group-mean centering of YEARS, COURSES, BLK, FRL, and ESOL; **p<.01

**Full model**. Six predictor variables, GROUP, YEARS, COURSES, BLK, FRL, and ESOL were included in this model as predictors of the intercept. GROUP shows that there was no difference in ESK at baseline for treatment and control teachers. Fixed effects show that none of the predictors were statistically significant (research question 2). Residual variance remains, however, in the intercept for ESK across teachers. Other school level predictors could be considered in future models.
Teachers in this sample showed strong negative perceptions of HSTA and a more positive perception of science standards that did not significantly change over the three years of the intervention. A SCT lens was used to help explain the cause of these results. This interpretation is presented in the following chapter.
CHAPTER 6

Discussion and Implications

Key findings with regard to each of the four research questions posed in the study are highlighted. Then, interpretations of results according to SCT are presented. Next, implications for PD and policy are presented. Finally, contributions and limitations are discussed.

Discussion

This study examined elementary science teachers’ perceptions of HSTA in a large, urban school district prior to any intervention (research question 1). At baseline, teachers in both the treatment and control groups viewed science standards more positively, but the state’s accountability policy in science negatively. The high-stakes science test was strongly perceived to impact teaching practices negatively by promoting didactic teaching practices such as drill and practice and prescribing instruction time to teaching to the test. Teachers also reported that they did not believe that the high-stakes science test scores were an accurate measure of ELLs’ science knowledge. These results are consistent with themes identified in the literature which found that teachers viewed standards positively and HSTA negatively (Abrams, Pedulla, & Madaus, 2003; Barksdale-Ladd & Thomas, 2000) and that HSTA impacts teaching practices negatively (Aronson, 2007; Coble, 2006; Hamilton et al., 2007).

This study also addressed whether teacher professional variables and school contextual factors predicted existing perceptions at baseline (research question 2). The number of science courses a teacher took and percent of Black students, low SES
students, and ELLs in the school did not predict at baseline any of the teachers’

perceptions of HSTA tested. Years of teaching was a significant predictor at baseline

only of the perception that science standards provided good guidelines. The higher the

number of years of teaching reported, the more strongly a teacher perceived that science

standards provided good guidelines. This is contrary to Stuart Hammer (2004) who found

that attitudes about accountability and standards-based reform did not vary by years of

teaching.

This study went on to examine whether science teachers’ perceptions of HSTA in

the treatment group changed compared to the control group (research questions 3a and

3b). There was no significant difference in any of the teacher perceptions of HSTA

examined for teachers in the treatment group compared to teachers in the control group
during this three-year intervention.

Finally, this study proposed to answer whether teacher professional variables and

school contextual factors predicted elementary science teachers’ perceptions of HSTA in

the treatment group compared to the control group (research question 4a and 4b). This

question was not germane to the analysis because teachers’ perceptions in the treatment

group compared to the control group did not vary during the intervention; thus, this

question will not be addressed further.

**Interpretations of Results According to SCT**

According to SCT, the interaction between causal factors results in the production

of effects. To create a given effect at times requires or can include many factors. Specific

factors therefore can be associated with effects probabilistically. Reciprocal processes

also occur within the behavioral, environmental, and personal realms that may influence
the effects produced by causal factors. In this study, interactions between the two causal factors, personal (e.g., teachers’ perceptions) and environmental (e.g., the intervention, HSTA, reform-based science), could ultimately result in notable effects for student learning.

**Baseline trends (research question 1).** A variety of strong negative perceptions of HSTA were reported by both treatment and control teachers at baseline which included the impact of HSTA on teaching practices, the reasonableness of the state accountability policy, and the inaccurateness of measurements of ELLs’ science knowledge by the high-stakes science test (i.e., personal factors). The interaction of these negative perceptions with the influence from the state’s HSTA policy (i.e., environmental factors) is likely to result in teachers’ use of non-preferred teaching practices such as drill and practice and *teaching to the test* to raise test scores. Although the perception that standards provided good guidelines was seen more positively, the most frequent mean response was found between somewhat agree and somewhat disagree. This set of mainly negative perceptions suggests that views of HSTA may prove especially difficult to change.

There is an existing trend in accountability policy to shift consequences for low performance from students and schools to teachers. New systems of teacher evaluation include a measure of student assessment scores (i.e., value added models). From a SCT perspective, it is reasonable to anticipate that teachers would respond less collectively and more individually in an effort to protect themselves from accountability sanctions (National Research Council [NRC], 2011). As this trend gains acceptance, teachers would be more likely to possess negative perceptions of HSTA and be less motivated to implement reform-based science when perceived as an ineffective way to maximize test
scores. Placing sanctions for low performance more directly on teachers is likely to reinforce strong negative perceptions of HSTA and strengthen constraints on behavior.

Baseline variations by teacher professional variables and school contextual factors (research question 2). This study identified a variation in teachers’ perceptions of standards at baseline by years of teaching which had not been recorded in the literature. Teachers with higher numbers of years of teaching more strongly agreed that standards provided good guidelines. The Stuart Hammer study found no variation in teacher perceptions of standards-based reform, but differed from this study in two important ways: (1) it sampled middle school instead of elementary teachers and (2) it took place around the time the state test was piloted and introduced, but before science counted as part of accountability. In contrast, this study began three years after science was tested and results counted as part of the state accountability program.

Middle school science teachers teach only science as compared to elementary teachers who teach multiple content areas. Middle school teachers may not feel the added tension that elementary teachers feel from teaching multiple subjects where some count more toward accountability (i.e., mathematics and English language arts) than others or not at all (i.e., science). Initial stages of implementation of a new initiative such as the introduction of a new state science assessment is usually accompanied by a greater level of PD, planning, and even anticipation than after implementation has been ongoing. For teachers in this study who have been teaching longer their experiences may be more defining because they have been a part of the history of change with standards and HSTA and possess more personal experience with students for comparison. It should be kept in
mind in future research and PD that years of teaching may predict teachers’ perceptions of standards.

It is noteworthy that no other professional variable or school contextual factor tested predicted teachers’ perceptions of HSTA at baseline for both treatment and control teachers. From a SCT perspective, teachers teaching greater numbers of ELLs or students from lower SES families (i.e., environmental factors) would most likely result in more negative perceptions of HSTA (i.e., personal factors) than teachers teaching primarily dominant student groups and students from higher SES families because they are found to generally attain significantly higher test scores. Teachers’ perceptions at baseline were more uniform than different with no significant variance by student population (i.e., dominant or non-dominant), student poverty level (i.e., higher or lower SES), or teacher content knowledge (i.e., number of science courses taken). This solidarity suggests that similar HSTA pressures are felt by a wide range of teachers in this district.

**Change in perceptions in the treatment group compared to the control group (research question 3a and 3b).** It is remarkable that despite what was observed to be a faithful implementation of the three-year intervention, teachers in the treatment group continued to report strong negative perceptions of HSTA such as increased drill and practice and teach to the test with no significant difference from teachers in the control group. This is unexpected since the intervention had wide support and followed research-based guidelines for PD.

Empirical research supports that an effective way to teach reform-based science is with inquiry-based learning (Akkus, Gunnel, & Hand, 2007), but also showed that teaching practices that promote reform-based science are not generally implemented. One
could argue that targets such as the ones in this study have been the objective of most PD programs for the last 20 years. Adamson, Santau, and Lee (2013) in their study of our previous project on elementary science teachers’ instructional strategies found that during interviews immediately following classroom observations, teachers did not report the use of more sophisticated inquiry strategies over a three-year intervention aimed at promoting inquiry and support for ELLs. Other factors than ineffective PD, insufficient resources and support, and lack of teacher content knowledge might be contributing to maintaining strong negative beliefs of HSTA in this sample of teachers.

SCT not only proposes that personal and environmental factors determine each other but that “reciprocality does not mean symmetry in the strength of bidirectional influences” (Bandura, 1986, p. 24). Since teachers’ perceptions remained unchanged despite the intervention, the influence posed by the demands of HSTA (i.e., environmental factor) appears stronger than the influence impacted by the intervention on teachers’ perceptions and likely to be a stronger influence on behavioral factors than that exerted by teachers’ perceptions (i.e., personal factor). SCT also argues that “when environmental conditions exercise powerful constraints on behavior, they emerge as the overriding determinants” (Bandura, 1986, p. 24). Evidence that the demands of HSTA may be acting as an overriding determinant of behavior is supported (1) by no impact by the intervention on teachers’ perceptions of HSTA and (2) by the literature which reported a positive perception of standards-based reform that does not generally translate to teaching practices that support such beliefs. As an overriding determinant of behavior, the effect of HSTA is likely to be the use of teaching practices to maximize test scores, even if the practices are contrary to beliefs.
Teaching practices would be a result of a variety of factors. The interactions between teachers’ perception that science standards provided good guidelines (i.e., personal factor) and the intervention (i.e., environmental factor) are expected to result in the use of teaching practices that promote standards-based reform. There was, however, no significant improvement in teachers’ perception that standards provided good guidelines and teachers at baseline only somewhat agreed with this statement. In this case, SCT suggests that teachers are not very strong advocates of standards-based reform, possibly due to the stronger environmental influence of HSTA. Plank and Condliffe (2013) concluded that even if the urban district they examined “successfully recruits and trains teachers so that they can improve instructional support, the results of our study suggest that accountability pressures may undermine these efforts since they may unintentionally encourage educators to use more teacher-centered pedagogical style and do not reward higher-order thinking” (p. 1177).

It is recognized that different factors could be responsible for similar effects. Not just HSTA, but also limited teacher content knowledge (i.e., personal factor) could result in the use of didactic teaching practices (i.e., behavioral factor). Results from examining measures of teacher content knowledge for this sample over the first two years of the intervention, however, showed that teacher content knowledge significantly increased over this time period for teachers in the treatment group compared to teachers in the control group. In addition, analysis of the impact of this intervention on student science proficiency (proficient versus not proficient) revealed that there was not a statistically significant difference in Year 1; the difference was statistically significant in Year 2 with students of teachers in the treatment group having a 28% higher odds of being proficient.
as compared to those in the control group in Year 2; and a statistically significant difference in Year 3 with students of teachers in the treatment group having a 36% higher odds of being proficient as compared to those in the control group. These results, taken together with results of unchanging and negative perceptions of HSTA from this study, suggest that teachers are acquiring skills they may not be using and students may be learning how to better take a test than developing scientific literacy as a result of the demands of HSTA.

Lotter, Harwood, and Bonner (2007) found that teachers’ perceptions strongly influence how teachers implement strategies learned in PD. Two occurrences at the end of Year 1 that might have enhanced negative perceptions of HSTA during this intervention included: (1) the adoption of more rigorous state standards and high-stakes assessments that teachers were informed were expected to result in significant drops in student test scores and (2) at that same time, exemption from participation in state tests for ELLs was changed from two years to one. The intervention addressed both changes by revising the curriculum and providing PD and SSS with a focus on strategies to navigate the new standards and assessments, especially for ELLs. At the end of Year 2, treatment teachers had better skills which they could have used to *teach to the test* as a result of the increased demands of HSTA. And increased test scores could have resulted in treatment teachers being more willing to do so. The lack of amelioration of teachers’ perceptions of HSTA supports this scenario. Future research should examine the impact of the intervention on teaching practices to corroborate this implication.
Implications for PD and Policy

**Implications for PD.** Teachers’ perceptions help determine teaching practices and this is critical because extensive research indicates that teaching practices is a key component for improving science achievement (Darling-Hammond, 2000). Interpretation of results using SCT highlight that key environmental factors, such as HSTA interact with teachers’ perceptions and likely influence if and how perceptions will be translated into behaviors. PD developed and implemented with a focus on one side of the SCT triangle (i.e., environmental and behavioral) may not be enough to effect desired change (i.e., behavior that promotes more meaningful student learning). Copur-Gencturk, Hug, and Lubienski (2014) found in their study of instructional practices of K-8 in-service teachers that:

- laying out a new vision for science teaching and learning and tying professional development to existing standards and school curricula are not enough. The current pressure to prepare for state tests that emphasize reading and mathematics and the (at least perceived) misalignment between high-stakes testing and reform-oriented teaching made at least some teachers feel they had to attend to breadth at the expense of depth in their attempts to cover the science curriculum in the limited time available. (p. 242)

Results from this study suggest that the interactions at the base of the SCT triangle (i.e., personal and environmental or teachers’ perceptions and intervention, reform based science, and HSTA) should be taken into consideration as future PD is designed and implemented. Because of the dynamic interplay between environmental and personal factors, those involved with the development and implementation of HSTA
policy and reform-based science should also take into consideration teachers’ perceptions and how these perceptions are likely to be translated into behaviors.

**Implications for policy.** Research shows that the implementation of reform efforts is usually tied to significant change in teachers’ beliefs (Levitt, 2001). This study found two significant results (1) teachers’ perceptions of HSTA were strongly negative before the intervention and (2) after a research-based curricular and PD intervention that was found to improve teacher content knowledge and student science proficiency, teachers’ perceptions remained unchanged and generally uniform across various professional variables and school contextual factors tested. This is especially salient as the NGSS were recently released and are introduced. If and when the NGSS are considered for adoption, it will be necessary to include teachers at the discussion table and avoid a traditional top-down trajectory to achieve desired implementation of standards-based reform. Including teachers in discussion will be especially valuable if and when the time comes for the development of corresponding assessments and the use of student test scores. In policy development and teacher PD, identifying and taking into consideration not just teachers’ perceptions, but also the interactions between teachers’ perceptions and environmental factors, such as HSTA and reform-based science which SCT suggest influence how beliefs are translated into practice, would be more likely to result in desired teaching behaviors and student learning outcomes.

**Contributions to the Literature**

This study makes several important contributions to the literature. First, it reports on the impact of a curricular and professional development intervention on teachers’ perceptions of HSTA in science education where studies were not found.
Second, it offers a possible explanation of how this intervention expected to ameliorate teachers’ perceptions of HSTA showed teachers’ perceptions did not differ over time by using SCT to interpret results. This conceptual framework provides an explanation of how the influence of key environmental factors may overrule beliefs and constrain behavior. This framework provides a possible perspective for use in future research.

Third, this study fills in several other existing gaps in the literature. This is an empirical study that used conceptually grounded measures to systematically collect data on HSTA for a large sample of teachers where the literature is largely based on case studies and descriptive research. In a field where much current research has not relied upon systematic measures of teachers’ perceptions or large samples of participants, this study provides instruments and data sets for future research. Results also indicate how change in teachers’ perceptions of HSTA is predicted by professional background and school contextual factors where little research is available.

Limitations

The study has several limitations that could be addressed in future research. First, the questionnaire is under development and could be improved to include only scales. Individual items may not be sensitive enough measures to capture change that may be taking place. Second, the instruments were administered to the treatment and control groups under somewhat different conditions. Control teachers could complete the measures at a comparable time and off-site location. Third, additional time points of data collection would allow additional growth models to be examined and more complex
statistical models to be used. Finally, teachers may have additional nesting within school; however, the sample size is not large enough to support a three-level hierarchical model.
REFERENCES


Rodgers, P. (2006). *What goal is of most worth? The effects of the implementation of the Texas Assessment of Knowledge and Skills on elementary science teaching* (Doctoral dissertation). Available from ProQuest Dissertations and Theses database. (UMI No. 3219180)


Appendix A

Synthetic Review of High-Stakes Testing and Accountability in Elementary Science Education in Relation to Student Diversity and Equity

High-stakes testing and accountability is a topic of great debate primarily for two reasons: (1) because it is the method by which the efficacy of standards reform is being measured in U.S. schools and (2) because of the effects it can have on student learning. Central to science education reform is the idea that “all students deserve and must have the opportunity to become scientifically literate” (National Research Council [NRC], 1996, IX). As a result, high-stakes testing and accountability has been implemented to ensure that all students receive the same high quality education prescribed by ongoing standards-based reform efforts.

As the primary interface between students and instruction, teachers are instrumental in the enactment of curriculum and student performance on assessments. Teachers’ perceptions of high-stakes testing and accountability may affect their teaching practices, which in turn could have an effect on student learning. Researchers have long acknowledged the role of teachers’ perceptions in their teaching practices and the impact on student outcomes (Anderson, 2011; Good & Brophy 1987; Lortie, 1975). It has been reported that “attitudes and expectations influence the climate of our classrooms, shaping what we teach and how we teach it, and thus have an impact on the achievement of our students” (Graybill, 1997, p. 311). Additionally, impacting teachers’ perceptions of high-stakes testing and accountability is the fact that teaching and learning is taking place with an increasingly diverse U.S. student population (Buckley, 2010).
The purpose of this review is to identify elementary science teachers’ perceptions of high-stakes testing and accountability and how these perceptions may vary by teacher demographic and professional background variables. Special consideration was given to issues of student diversity and equity, specifically English Language Learners (ELLs), the fastest growing student group in the United States. As a result, from the large body of articles in the area of high-stakes testing and accountability, this review focused on the limited literature on science teachers’ perceptions of high-stakes testing and accountability. Due to the limited amount of literature on science teachers’ perceptions of high-stakes testing and accountability, this review at times draws on available studies of teachers’ perceptions of high-stakes testing and accountability in other content areas such as English language arts and mathematics. The literature included in this review was selected and organized to respond to the following issues:

1. Identifying teachers’, and more specifically science teachers’, perceptions of high-stakes testing and accountability.
2. Highlighting science teachers’ perceptions of high-stakes testing and accountability in relation to issues of student diversity and equity.
3. Recognizing whether science teachers’ perceptions of high-stakes testing and accountability vary by their demographic and professional background.

Defining High-Stakes Testing and Accountability

In the United States, two major science educational reform efforts have taken place (Duschl, 2008). The first reform occurred in the 1950s after the launch of Sputnik by the former Soviet Union. After this event, the United States concluded that the only way to put the country back in first place in the space program and maintain its military
advantage was by improving the science achievement of its populace. The National Science Foundation (NSF) was formed and the nation’s science curricula were reformed with the goal of producing little scientists. The second reform effort was sparked three decades later in 1983 by the publication of *A Nation at Risk: The Imperative for Educational Reform*. This report called for identifying practical recommendations to be implemented by educators in the classroom to improve student achievement in science because of a perceived crisis of underachievement in U.S. schools, which if not corrected would leave the United States with a populace unable to effectively compete internationally (National Commission on Excellence in Education, 1983).

As a standards movement has become the focus of U.S. education, more emphasis has been placed on test-based accountability systems as a means to measure whether all students are achieving proficiency in the standards. Assessment in the form of standardized testing that is used for the purpose of accountability is referred to as high-stakes testing (Resnick, 2000). Accountability indicates policies that mandate large-scale testing to measure attainment of state and/or federal goals. Ideally, accountability policies will motivate teachers, administrators, and others to work toward improving educational opportunities and student learning. Accountability systems usually reward schools for high performance and impose sanctions on schools that fail to move students in all identified subgroups toward proficiency. Standards, classroom instruction, and assessments must be aligned if the accountability system is to be effective. Assessments, no matter how well aligned with standards, can only measure a subset of the total content that could possibly be assessed (NRC, 2001).
The No Child Left Behind Act (NCLB), passed by Congress with bipartisan support and signed into law by President George H. Bush in January 2002, has made high-stakes testing and accountability an integral part of every U.S. classroom. As mandated by NCLB, by 2005-2006 reading and mathematics is tested annually in grades 3 through 8 and once in high school in all states. Beginning in 2007-2008, NCLB mandates that science is tested at least once in grades 3 through 5, 6 through 9, and 10 through 12, although Annual Yearly Progress (AYP) in science is not required to be measured and reported. Science is made part of school accountability at specific grade levels in some states and generally contributes less to the overall school performance than reading and mathematics. Disaggregated results for demographic subgroups of students by major ethnic/racial groups, economically disadvantaged students, limited English proficient (LEP) students, and students with disabilities are made publicly available annually.

A decade after the implementation of NCLB, it has become obvious that its mandate – by 2014 100% of U.S. students would be proficient in science and mathematics – would not be achieved. In September 2011, the U.S. Department of Education reported the following to states and districts:

While NCLB helped State and local educational agencies (SEAs and LEAs) shine a bright light on the achievement gap and increased accountability for groups of high-need students, it inadvertently encouraged some States to set low academic standards, failed to recognize or reward growth in student learning, and did little to elevate the teaching profession or recognize the most effective teachers. (Duncan, 2011)
As a result, a process was put in place for states and districts to request flexibility regarding specific requirements of NCLB. By August 2012, 32 states and the District of Columbia had been granted waivers. To be eligible to receive the waiver, states had to develop and implement plans to improve educational outcomes as outlined by the Obama Administration education agenda, such as using students’ standardized test scores as part of their teacher evaluation systems.

NCLB has generated substantial debate. While the effects of NCLB have been largely criticized, NCLB has also had positive impact. First, non-dominant groups are now part of the educational discourse because of NCLB (Cizek, 2001). Second, the Act has led to increased instruction time for subject areas, such as science, which previously were underemphasized (Diamond & Spillane, 2004; Rodgers, 2006). Increased science instruction is especially noted in years when high-stakes testing is administered. Third, following passage of NCLB there has been an increased rigor in science curriculum and a better alignment of curriculum with standards and assessment (Hamilton et al., 2007; Rentner et al., 2006).

Nevertheless, high-stakes testing and accountability continues to be a divisive topic. Cizek (2001) proposed that one reason why high-stakes tests are poorly accepted is that the people responsible for writing and administering them spend very little time informing teachers and the public about the benefits of the tests and the rationale behind them. Advocates see it as the way to motivate learners, inform teachers, and identify ineffective schools for rehabilitation. Critics suggest that unintended consequences are too negative and include a narrowing of the curriculum, increasing student anxiety and teacher stress, and perpetuating minority disparities.
Non-dominant communities such as low-income households, homes where students speak a language other than English, African Americans, and Hispanics or Latinos/as continue to be impacted differently by standardized testing. Significant testing gaps among diverse student groups continue to exist. For example, a Black/White gap and a Hispanic/White gap have been largely stable over the past decade (Jordan, 2010) and English-proficient students perform better on standardized tests than students speaking a language other than English at home. Despite having implemented reform-based science in U.S. classrooms since the early 1980s, equitable learning opportunities in science for all students remain elusive (Jordan, 2010; Lee & Buxton, 2008).

It is generally agreed that it is the responsibility of the U.S. school system to generate individuals possessing a range of scientific knowledge and ensure that our schools offer the opportunity to possess such knowledge to every student. In a high-stakes testing and accountability policy context, unpacking science teachers’ perceptions of high-stakes testing and accountability will highlight whether existing teachers’ perceptions are facilitating the acquisition of scientific knowledge or impeding it.

**Literature Review of Teachers’ Perceptions of High-Stakes Testing and Accountability**

The demands for accountability largely come from politicians and policymakers (Smith & Fey, 2000). Despite the understanding that teachers’ perceptions matter, teachers have not played an active decision-making role in the development of reform policies (Barksdale-Ladd & Thomas, 2000; Crocco & Costigan, 2007). There has, however, been a call in the literature that “a clear understanding of educators’ perceptions of the impacts of current test-based accountability policies should guide the development
and implementation of the next generation of national science standards and subsequent
large-scale assessments” (Anderson, 2011, p. 1).

**Teachers’ Perceptions of High-Stakes Testing and Accountability**

Even before high-stakes testing was the educational norm in the United States, Jackson (1968/1990) reported in his now recognized classic work, *Life in Classrooms*, that elementary teachers showed “a general distrust of tests” (1990, p. 124). Teachers saw a very limited number of useful instruments and believed practices to administer tests were poor. These reports were based on standardized test results not being reported to the teacher consistently or the data arriving too late in the school year to do much good. Teachers further believed that children behave atypically on tests and that performance on achievement tests is more a reflection of native ability than of teaching effectiveness (Jackson, 1990).

As the standards movement took hold and high-stakes testing and accountability began being widely implemented, substantial debate ensued on the effects of high-stakes testing and accountability. Although the majority of unintended consequences of high-stakes testing and accountability are negative, researchers identified some positive effects. Cizek (2001) lists nine benefits that resulted from high-stakes testing: (1) increased professional development, (2) accommodation for students with special needs, (3) educator knowledge about testing, (4) collection and use of data in making decisions, (5) increased educational options for students, (6) the beginnings of accountability systems, (7) improved educators’ intimacy with their disciplines, (8) improved quality of texts, and (9) increased student learning. Massachusetts saw average student scores on national tests shift to be the top in the nation in the years following the institution of high-
stakes testing (Reville, 2004). In North Carolina where more attention was given to fourth and fifth grade students with learning disabilities since their inclusion in high-stakes testing, a marked increase was noted in reading ability (Schulte, Villwock, Whichard, & Stallings, 2001). Elementary teachers reported an increase in the use of interdisciplinary lessons (Smith & Kovacs, 2011). Bishop (2000) showed that curriculum-based external exit exam systems, such as those used in New York and Canada, caused students to outperform students from areas that did not use such tests, which indicates that the students had learned more possibly because:

> [S]chools in exit exam provinces scheduled more hours of math and science instruction, assigned more homework, had better science labs, and were significantly more likely to use specialist teachers for math and science and more likely to hire math and science teachers who had studied the subject in college.

(Bishop, 2000, p. 212)

While student standardized test scores often go up after the implementation of high-stakes testing, other indications of student learning tend to go down, which suggests that the only learning that is increasing is how to pass a specific test. Amrein and Berliner (2003) found that in 18 states the testing requirements for graduation ACT and AP scores went down and SAT and NAEP scores showed no significant change. Even though teachers generally agreed that standards provided good guidelines, they had serious concerns about the negative effects of high-stakes testing on teaching practices and teachers’ and students’ well-being (Abrams, Pedulla, & Madaus, 2003). There is general agreement that teachers tend to teach to the test, especially in a high-stakes environment (Daggett, 2000; Hamilton et al., 2007; Herman, 2005; Schroeder, 2003).
Impact on teaching practices. Teachers reported that high-stakes testing impacted practice on a daily basis (Lewis, 2003). Smith and Kovacs (2011) found that “elementary teachers were the most likely to say that they have little time to teach things not on state-mandated tests” (p. 213). If the state mandated test follows a multiple-choice format, classroom assignments will also follow this format (Abrams et al., 2003). Noddings (2004) recognized that since testing critical thinking skills is difficult, especially in a multiple-choice format, teachers tended to overlook using instructional practices that promoted higher-level thinking. This choice is also possible in the absence of high-stakes testing and accountability. A high-stakes testing and accountability context could be perceived as the justification and serve to increase the number of teachers overlooking the use of instructional practices that promote higher-level thinking. Moreover, significant class time was devoted to test preparation activities (Abrams et al., 2003). Four in ten teachers surveyed in one study stated that test scores can be increased without learning occurring (Schroeder, 2003). Teachers also reported “focusing more on students near the proficient cut score (i.e., ‘bubble kids’) and expressed concerns about the negative effects of the accountability requirements on the learning opportunities given to high-achieving students” (Hamilton et al., 2007, p. XIX).

Teachers believed that high-stakes testing has led them to teach in ways that contradicted their idea of effective teaching practices (Barksdale-Ladd & Thomas, 2000; Schroeder, 2003). The perception is that drill and practice lessons are increasing and inquiry lessons are decreasing (Settlage & Meadows, 2002). Teachers at schools where students traditionally do well on high-stakes tests are seen to have significantly more freedom to employ best practices, because these schools are less affected by the
repercussions of low test scores (Eick, 2002). This can lead to a widening achievement gap, as the students in high-achieving schools have teachers who are encouraged to teach using more effective teaching practices. The assessed curriculum is viewed as resulting in limiting teaching practices (Crocco & Costigan, 2007). There was also general agreement among teachers that there was less emphasis on science in the elementary grades where the emphasis is on the tested subjects of English language arts and mathematics, leaving little to no time to teach science (Crocco & Costigan, 2007; Diamond & Spillane, 2004; Jennings & Rentner, 2006). High-stakes testing is further seen to lead to a narrowing of the curriculum (Crocco & Costigan, 2007; Diamond & Spillane, 2004). Teachers of students who have already finished high-stakes testing, for example seniors in high school, claimed that high-stakes testing does not affect their teaching practices (Eick, 2002).

Some schools have developed curricula that are specifically designed to integrate learning goals with test-preparation (Eick, 2002). In Mississippi and Tennessee, high-stakes testing is reported to have led teachers to combine direct instruction with cooperative group learning, which is considered by some as a better method for students to learn (Vogler & Burton, 2010). Allen (2005) points out that some test-taking strategies, such as having students exposed to a variety of reading passages while teaching them to use context clues, also results in content learning.

**Impact on teachers’ sense of professional well-being.** High-stakes testing has impacts on teacher practice that result in teaching methods that are not aligned with existing theories of learning on which reforms are based. Current understanding of learning has resulted in a much different role for the teacher, who is no longer a sole
holder and dispenser of knowledge, but a facilitator of meaning-making. Implementation of standards is recognized as a multiphase process that includes: (1) student-generated questions, (2) student-designed and constructed models, (3) group discourse which will refine the model and ideas, and (4) meaningful conclusions supported by evidence. The science teacher uses problem solving, cooperative learning, and hands-on activities as learning strategies, while providing appropriately challenging activities, allowing students to take authority over their learning, making sure that their work can be commented on by other students and teachers, promoting students to actively listen and respond, and ensuring access to necessary resources.

The tension between the role teachers perceive is required of them in a climate of high-stakes testing and accountability and the role they believe they should be pursuing in the classroom could translate into teacher dissatisfaction. While happy teachers would be a boon, it is not being suggested that teacher satisfaction should be the ultimate goal of science education reform. The research literature is rich on the impact of a demoralized workforce on organizational productivity. The effect of conflicting tensions on teachers could be negative perceptions and less effective practices. For this reason the impact of high-stakes testing and accountability on teachers’ sense of well-being were highlighted in this review.

Teachers reported negative effects on their well-being as a result of high-stakes testing and accountability (Barksdale-Ladd & Thomas, 2000; Mintrop, 2004). Teachers sensed pressure from policy mandated testing, felt a lack of agency and control, and believed that testing is forcing them to choose between the creative teaching practices that can lead to exceptional student learning and the drill work that will make the average
student successful on the high-stakes test (Barksdale-Ladd & Thomas, 2000; Sutton, 2004). Crocco and Costigan (2007) reported that as a result of high-stakes testing teachers “felt the lack of time available for developing projects or digging more deeply into subject matter relevant to students damaged their efficacy” (p. 522). Teachers reported a frenzied pace to be able to cover the extensive material included in high-stakes assessments (Crocco & Costigan, 2007). Abrams et al. (2003) found that teachers wanted to be reassigned from tested grades to grades where students were not assessed. Nevertheless, Barksdale-Ladd and Thomas (2000) found that “no teacher argued for a lack of standards or for dismissing assessment” (p. 395).

As high-stakes testing and accountability policy has become more entrenched, a larger number of studies have focused on empirically examining science teachers’ perceptions on the topic as it applies to science education. Anderson (2011), in his synthesis of studies on science education and test-based accountability policies, identified 35 such studies of which the majority were identified as using ethnographic or case study methodology. From these varied studies, which included work in elementary, middle, and high schools from across the United States and other countries, some broadly held science teachers’ perceptions on high-stakes testing and accountability have emerged. For example, science teachers perceived standards positively but the impact of high-stakes testing and accountability on teaching practices and teachers’ sense of well-being negatively.

**Impact on science teachers’ teaching practices.** Science teachers credited high-stakes testing with increasing curricular rigor and better aligning curriculum with standards and assessments (Anderson, 2011; Hamilton et al., 2007). Vogler (2002) found
that a stratified random sample of teachers teaching at least one section of tenth grade English, mathematics, or science in a public high school in Massachusetts self-reported a greatly increased use of inquiry-based lessons when high-stakes test scores first became public. More specifically, 25% of teachers reported increased use of lab equipment, 31% reported increased use of modeling, and 42% reported increased use of computers. Additionally, 12% of teachers reported decreased use of textbook-based assignments and 36% reported decreased lectures. Data in this study were obtained from a 54-question survey instrument and analyzed using frequency tables, cross tabulations, and grouping of comments according to content. Vogler (2002) does not provide details as to how strategies were implemented, only whether or not the teachers felt they used them. The teachers most likely to change their teaching practices were those who had been teaching for 13-19 years, while both more and less experienced teachers reported less change. The most common reason teachers gave for changing their teaching practices was to help students pass the test so the students could graduate.

Contradictory to the results presented by Vogler (2002), other studies found that teachers reported a decrease in the use of inquiry lessons and an increase in fact-based learning (Coble, 2006; Jenkins, 2000; Settlage & Meadows, 2002). Science teachers also believed that teachers of non-tested courses used more inquiry-based curricula than teachers of courses that are being assessed as part of accountability policies (Coble, 2006). Moreover, science teachers believed that high-stakes testing led to a focus on teaching to the test (Font-Rivera, 2003; Hamilton et al., 2007) and that science teachers devoted significant class time to teaching test-taking skills (Font-Rivera, 2003; Hamilton et al., 2007). The teachers in Font-Rivera’s study (2003) reported that the state test did
not accurately portray learning. Diamond and Spillane (2004) noted that, in schools on
probation for not having met testing requirements, instructional time dedicated to science
decreased or was even eliminated. Science teachers identified in grades where high-
stakes testing was not taking place, however, reported that they perceived minimal impact
of high-stakes testing (Wideen, O’Shea, Pye, & Ivany, 1997).

**Impact on science teachers’ sense of professional well-being.** Science teachers
felt less professional satisfaction as a result of high-stakes testing and accountability
(Anderson, 2011; Jenkins 2000). They perceived that high-stakes testing and
accountability policies resulted in a decreased sense of autonomy (Mintrop, 2004), an
increased level of stress and decreased morale (Aronson, 2007; Rentner et al., 2006), and
pressure to increase test scores (Font-Rivera, 2003). In addition, they agreed that high-
stakes testing led them to teach in ways that compromised their view of effective teaching
(Aronson, 2007) or did not allow them to explore topics of student interest (Galton,
2002).

The scholarly work on science teachers’ perceptions of high-stakes testing and
accountability that was reviewed appeared in peer reviewed journals or was an accepted
dissertation that included empirical data. Few studies, however, used stratified random
samples of teachers (Anderson, 2011). Some studies included very few numbers of
schools and/or teachers in their sample. This review also highlighted that existing work
has been done nearly twice as often with middle and high school teachers as with
elementary teachers.
Teachers’ Perceptions of High-Stakes Testing and Accountability and Issues of Language Diversity and Equity

As previously mentioned, one of the most significant outcomes of high-stakes testing and accountability has been the increased focus on achievement gaps and the learning needs of marginalized student groups. Issues of multicultural education are more salient as the U.S. student population grows more diverse. According to the U.S. Census, it is projected that racial and ethnic minorities will make up half of individuals under 19 years of age by 2022 (U.S. Census Bureau, 2012). One of the most significant demographic shifts currently affecting U.S. education is the growing number of ELLs (Callahan, 2005), which is continuing to grow annually (National Center for Education Statistics, 2012). Twenty-one percent of children ages 5-17 spoke a language other than English as their primary language at home and spoke English with difficulty. In 2009 over 5 million, or 11% of U.S. public school students in K-12, were identified as ELLs. Taking into account the needs of ELLs is important as a greater number of teachers will be tasked with preparing them for successful performance on high-stakes testing. Under existing accountability policies, ELLs, with minimal exceptions, are expected to take part in all assessments and have their scores count as part of various accountability measures. The stakes are high not only for schools and teachers, but for individual students as well. Student scores are routinely used as a deciding factor to deny a student’s promotion or graduation.

Closely intertwined with the reality of diversity is the issue of equity. In the literature, scholars use the term *equity* in varied ways. Uses of equity include: equity as physical access, equity as inclusion, equity as multiculturalism or diversity, and equity as
special services (Powell, Boethel, & Southwest Educational Development Lab, 1994). To have equity with respect to special services includes providing special classes, mentoring opportunities or even social services to special groups as needed. The special groups discussed in science education typically include low-income households, ELLs or limited English proficient (LEP), African-Americans, Hispanics or Latinos/as, and Native Americans. Use of broad terms such as minority groups is also common in the literature making it unclear which specific group is being considered. This review addresses ELLs specifically and interprets equity in the broader term where “equitable learning opportunities occur when formal schooling (a) values and respects the knowledge and experiences students bring from their home and community environments, (b) articulates such knowledge and experiences with academic disciplines, and (c) offers educational resources and funding to support all students’ learning” (Lee & Buxton, 2008, p. 124).

**Bilingualism – benefit or disadvantage.** The literature is substantial when it comes to teachers’ interactions and attitudes toward non-dominant student groups and teachers have been found to have strong beliefs about non-dominant students. Teachers have cultural biases toward low SES students (Payne, 1994); their expectations and speech vary with students’ ethnic background (Tenenbaum & Ruck, 2007); and they have different educational aspirations for students with different ethnicity/race (Cheng & Starks, 2002). Additionally, teacher beliefs are influenced by track placement (Callahan, 2005) and teachers have lower expectations for students in low tracks (Oakes, 1990).

In the United States, students who are non-native speakers and/or are in the process of acquiring the language of school as their second language have historically been perceived as deficient. This has been in large part due to the intransigent view that
the consequences of knowing more than one language have negative consequences on cognitive development. This has remained despite the fact that research to the contrary has been available since the mid-1930s (Barac & Bialystok, 2011). As the field of second language acquisition has progressed asking better defined questions, using more rigorous methodology, and undertaking a larger number of studies, especially since the 1990s, confirming research has mounted that “bilingualism turns out to be an experience that benefits many aspects of children’s development” (Barac & Bialystok, 2011, p. 36). Benefits have been noted in metalinguistic awareness, cognitive development, and multi-tasking where bilingual children have been found to perform better than monolingual children (Bialystok, Barac, Blaye, & Poulin-Dubois, 2010).

Inquiry into bilingual language phenomena such as code-switching provides further supporting evidence that bilingualism is not a deficiency but a benefit. Code-switching is the act of changing from one language to another during a conversation, in some cases even in mid-sentence, between bilinguals who share the same language background (Wei, 2007). Code-switching is a practice observed in ELLs by teachers in schools. On the surface, code-switching appears, and therefore can be perceived, to be a natural response for an individual who is simply unable to communicate in one language and is turning to the other language in which he/she can better communicate.

Research, however, has shown that code-switching is in fact not a result of a language deficiency, but only possible because the users possess language competency in more than one language (Poplack, 1980; Silva-Corvalán, 2007; Zentella, 1997). Empirical evidence shows that there are rules that govern code-switching, that it is used for important purposes and specific functions, and that only the most competent bilingual
students use code-switching. This understanding challenges the perception that bilingualism hinders intellectual development, and today many linguists recognize that bilingualism benefits aspects of an individual’s communicative, cognitive and cultural spheres (Wei, 2007).

While current research on bilingualism does not support a deficit model, our school culture still does. This may be in part due to the fact that teacher preparation and professional development rarely highlight research and resulting theories of bilingualism or second language acquisition. Palmer and Martinez (2013) believe that “developing more robust understanding of bilingualism and the interactional dynamics of bilingual contexts will help teachers learn to better engage bilingual learners” (p. 273). They further assert that “all of this matters, we argue, because they [teachers] are uniquely situated as powerful actors within bilingual classroom contexts” (p. 289). This is ever more salient since “the language and literacy demands that undergird the new standards clearly present challenges for all students, but particularly for students who are still in the process of learning the language of instruction” (Bunch, 2013, p. 299).

**Disproportionality in special education.** Placement and issues of disproportionality in special education are very complex. Nevertheless, disproportionate representation of non-dominant students in special education has been well documented (Artiles, Kozleski, Trent, Osher, & Ortiz, 2010). The seminal works of Dunn (1968) and Deno (1970) first highlighted the overrepresentation of non-dominant students in special education. Artiles and Trent (1994) and again Artiles et al. (2010) argue that overrepresentation is a critical issue that still remains a problem. Additionally, the most recent NRC report confirmed that “learners from historically underserved groups are
disproportionately represented in high-incidence disability categories (mild mental retardation [MMR], learning disability [LD], and emotional/behavioral disorder [EBD])” (Artiles et al., 2010, p. 280).

The data on disproportionality for ELLs with special needs is scarce. How much we clearly understand is limited. Some of what we do know is that our understanding is complicated by the fact that the term minority is treated as a homogenous group when in fact it is composed by a variety of heterogeneous subgroups, each affected by different factors and defined by a variety of different social constructs – race, ethnicity, nationality, SES, etc. Moreover, ELLs are also a heterogeneous population. For example, ELLs are comprised of individuals who speak different languages, have different ethnic backgrounds, and come with different immigration status and variety of years in the United States. Further complicating matters is the fact that precise information about students’ levels of language proficiency in English and in their native language is lacking, since adequate measures of language proficiency do not exist and states have kept incomplete data (Klingner, Artiles, & Barletta, 2006).

From the available data, Klingner et al. (2006) reported that ELLs were overrepresented in special education and more specifically in LD categories. Fifty-six percent of ELLs identified with special needs are designated LD with reading difficulties, followed by 24% designated with speech-language impairment. A disproportionate number of Latino students are designated LD (Artiles et al., 2010). This is significant because compared to ELLs without disabilities, ELLs designated with disabilities receive fewer language support services, tend to be instructed only in English, are segregated for
services, and have instructional programs that are not as aligned with state performance standards (Klingner et al., 2006).

The identification of high-incidence disabilities relies on a highly subjective process. Since no clear biological or genetic etiology is present, diagnosis could be misguided by personal perceptions and biases of the assessor. Overrepresentation of non-dominant groups in special education may be a result of misguided teachers’ perceptions of non-dominant students and inadequate preparation to adequately assess multilingual students (Artiles et al., 2010). More placements could translate to more services and ultimately better outcomes for those placed. Unfortunately, a special education designation is not a guarantee to a better educational future. Special education students on average perform below general education students on the National Assessment of Education Progress (NAEP) assessments (Cortiella, 2007).

Klingner et al. (2006) suggest that underrepresentation of ELLs in special education should also be a concern. Some general education teachers have not received adequate preparation to be able to determine if the causes of reading difficulties of an ELL stem from the second language acquisition process or are a result of a LD. Miss-designations could help explain why ELLs in special education and more specifically identified as LD are underperforming on standardized testing.

**Standardized testing bias.** The concept and intent of testing is worthwhile and necessary. “Assessments of school learning provide information to help educators, administrators, policy makers, students, parents, and researchers judge the state of student learning and make decisions about implications and actions” (Pellegrino, 2012, p. 832). But because it is measuring an intangible mental representation or process, an assessment
is also recognized as a “tool designed to observe students’ behavior and produce data that can be used to draw reasonable inferences about what students know” (Pellegrino, 2012, p. 833). If the data being produced are not being used to draw *reasonable inferences* for the individuals being assessed, then the resulting consequences are invalid and undeserved. Pellegrino (2012) reports that, “many assessments are insufficient on a number of dimensions including representation of the cognitive constructs and content to be covered and uncertainty about the scope of the inferences that can be drawn from task performance” (p. 835).

State mandated science assessments cannot test every standard primarily for two reasons: (1) use of mainly a multiple-choice format and (2) limitations to monitor the length of the test. The increasing reliance on a multiple-choice format is expected. In a poor economic climate, and since the passing of the NCLB Act which requires more testing across a larger number of grades and fast reporting in areas such as AYP, there has been a greater use of multiple-choice tests across all states (Beatty, 2010).

Unfortunately, a multiple-choice format is not able to assess some forms of meaningful learning.

NCLB further dictates that the test scores of ELLs be included when measuring educational progress. Test scores for ELLs are excluded only during their initial years in ESOL/ESL programs. This is usually two years, but can vary. The law further requires that ELLs be provided any necessary accommodations when taking high-stakes tests. Accommodations vary from state to state and include a variety of adjustments in terms of timing and scheduling of the test, the place in which testing takes place, the use of bilingual dictionaries, and the presentation of test materials or test directions.
It has been suggested, however, that interpretations of science assessment scores, specifically for ELLs, lend themselves to being invalid (Noble et al., 2012; Penfield & Lee, 2010). Moreover, significant performance gaps continue to exist for non-dominant groups (Jordan, 2010; Lee & Buxton, 2008). Diamond and Spillane (2004) and Eick (2002) argue that the perpetuation of such educational gaps, despite reform efforts, may be explained by different responses to accountability policies by low- and high-achieving schools. Designation as a low-achieving school often results in schools focusing on being removed from such a list in an effort to avoid possible sanctions down the road. Sanctions applied to schools not in compliance with NCLB include, but are not limited to, adopting and implementing school improvement plans, having their accreditation denied, and even facing possible closures. Test scores are also being used as a part of teacher evaluation systems to determine such things as teacher retention, promotion, and merit pay. Low-achieving schools tend to rely on practices targeted at increasing test scores, even if such practices are markedly at odds with promoting learning for understanding and developing critical thinking skills. In contrast, high-achieving schools tend to focus on reward and encouragement of continued instructional improvement by following preferred practices independent of accountability policies.

High-stakes testing and accountability policies may be having an opposite and unintended consequence in our schools. High-stakes testing and accountability may continue to reproduce social inequities and limit academic opportunities in an increasingly diverse educational system. As a result, calls such as the following are being made:
In the world of standards-based reform, which seeks to align key components such as curriculum, instruction, teacher capacity, assessment, and system support through explicit and coherent standards, issues of ELLs need to be kept in mind and infused into the systemic changes. (Hakuta, 2010, p. 167)

Understanding the nature of bilingualism itself is complicating researchers’ understanding of why ELLs continue to underperform on standardized tests. In line with the theory of a unitary language system where representations for all different languages are stored in an interconnected system, Valdés (1994) proposes that the verbal ability of bilinguals being evaluated in a setting in which only one language code is allowable would appear inferior to what the student’s ability actually is. The bilingual does not have the knowledge base to be assessed with a monolingual’s test. Therefore, “the bilingual test taker cannot perform like a monolingual, and the monolingual test cannot measure in the other language” (Valdés, 1994, p. 2).

Valdés (1994) also makes the distinction that a majority of U.S. bilinguals, who are immigrants, are “circumstantial” instead of “elective” bilinguals. Circumstantial bilinguals must learn a new language because they emigrated from their country, whereas elective bilinguals choose to learn a new language without other compelling factors. Her resulting argument is that circumstantial bilinguals cannot be appropriately tested with instruments designed for use with elective bilinguals. If the measures cannot assess the achievement of (circumstantial) bilingual students, we do not have any actual information on ELL achievement. This puts into question the validity of the high-stakes decisions based on the scores on these mandated assessments. Educational policies that impact
ELLs, however, are often enacted without basis on existing second language acquisition research (Dixon et al., 2012).

Psychometrics has also been working to unpack issues of testing bias for ELLs. Noble et al. (2012) report that “statistical studies have shown that large-scale tests show lower reliability for ELLs than for non-ELLs, lower correlation with other ability measures (Abedi, 2002), and that test items with greater linguistic complexity tend to show greater bias against ELLs (Martiniello, 2009; Wolf & Leon, 2009)” (p. 781). Research has also been focused on issues of validity to determine whether high-stakes tests are actually measuring the knowledge they are intended to measure specifically for non-dominant groups. When investigating students’ response processes when answering science test items, Noble et al. (2012) found that “students from low-income households and English Language Learners were more likely than middle-class native English speakers to answer incorrectly despite demonstrating knowledge of the targeted science content for the items” (p. 778). The measurement field’s work lends support to the hypothesis that standardized test scores do not accurately reflect students’ science knowledge as intended by high-stakes assessments and possibly explain test score gaps for ELLs.

**Teachers’ perceptions of high-stakes testing and accountability for ELLs.**

One of the most significant outcomes of high-stakes testing and accountability has been the increased focus on achievement gaps and the learning needs of marginalized student groups. Issues of multicultural education are more salient as the U.S. student population grows more diverse. According to the U.S. Census, it is projected that racial and ethnic minorities will make up half of all individuals under 19 years of age by 2022 (U.S.
Census Bureau, 2012). One of the most significant demographic shifts currently affecting U.S. education is the growing number of ELLs (National Center for Education Statistics, 2012). Among children between 5-17 years old, 21% spoke a language other than English as their primary language at home and spoke English with difficulty (National Center for Education Statistics, 2012). In 2009 over 5 million, or 11% of U.S. public school students in K-12, were identified as ELLs (National Clearinghouse for English Language Acquisition, 2011). Addressing the needs of ELLs is important as a greater number of teachers are tasked with preparing ELLs for successful performance on high-stakes testing.

Under existing accountability policies, ELLs are expected to take part in all assessments and have their scores count as part of various accountability measures with minimal exceptions. Allowances from assessments and accountability are generally granted for the lowest levels of LEP students for the first two years or in some cases only one year, of full-time attendance in a U.S. school. The stakes are high not only for schools and teachers, but for individual students as well. Students’ scores are routinely used in determining their placement in advanced courses, as well as promotion or graduation.

There are particularly effective teaching methods that come from research for science education broadly and for ELLs specifically (Lee & Buxton, 2008). In supporting ELLs to learn science content and acquire second language simultaneously, effective practices include: (1) multiple modes of communication and representation such as verbal, gestural, written, and graphic; (2) extensive graphic materials such as Venn diagrams, drawings, tables, and graphs; (3) introduction of vocabulary at the beginning of
units with multiple opportunities for practice in a variety of settings; and (4) extensive opportunities for hands-on and inquiry-based learning (Lee & Buxton, 2008). Inquiry-based learning with explicit attention to vocabulary has been repeatedly found to result in higher performance on district-wide science assessments (Lara-Alecio et al., 2012). The literature is fairly consistent that mainstream teachers need specialized knowledge and preparation to work with ELLs (Bunch, 2013; Lucas, 2011).

Many elementary science teachers, however, felt unprepared to teach science (Banilower, 2013) and to address student diversity (National Center for Education Statistics, 2012). Many teachers believed that the standards implemented were often not appropriate for the diverse students they were assigned to teach (Barksdale-Ladd & Thomas, 2000). Bilingual education teachers believed that largely because of high-stakes testing they were ignored when it came to contributing to the best course of action to teach ELLs (Palmer & Garcia, 2000). Teachers of ELLs must account for dissimilar prior knowledge, varied cultural background resulting in different beliefs and misconceptions, and proficiency in home languages different from that being used in the classroom and on high-stakes assessments. As ELLs are pressured to learn English, the language of the test, there is concern that the students will fall further behind in academic subjects such as science without time to incorporate the scaffolding provided by their native language.

Science teachers viewed high-stakes testing and accountability as making it especially difficult to offer non-dominant student groups, such as ELLs, the opportunities they are entitled from the education system (Penfield & Lee, 2010). Teachers were concerned that students pulled out of content courses for ESOL instruction were not receiving the science instruction they were to be tested on (Shaver, Cuevas, Lee,
Avalos, 2007). Wright and Choi (2006) found that teachers viewed state-mandated tests as not being equitable for ELLs:

> In terms of accountability, while the overwhelming majority of teachers support the general principle, they believe that high-stakes tests are inappropriate for ELLs and participants provided evidence that the focus on testing is leading to instruction practices for ELLs which fail to meet their unique linguistic and academic needs. (p. 2)

**Teachers’ Perceptions of High-Stakes Testing and Accountability by Demographic and Professional Variables**

Identifying whether teachers’ perceptions vary by professional and demographic background is important because “people in a similar line of work are likely to share at least some common thoughts and feelings about that work . . . and . . . it is essential to know their nature if we are to grasp the ethos of an occupation” (Lortie, 1975, p. 162). Also a strong argument exists for the idea that teachers’ perceptions matter in teaching and learning (Bandura, 1986; Graybill, 1997; Pajares, 1992) and more specifically that they matter in the area of testing and accountability (Cimbricz, 2002). A review of the literature concluded that:

> a relationship between state-mandated testing and teacher beliefs and practice is consistently confirmed. State-mandated tests do matter and do influence what teachers say and do in their classrooms. But while there is overall agreement that a relationship between the two does exist, the nature of that relationship is neither simple nor easy and requires further elucidation. (Cimbricz, 2002, p. 4)
Few studies examining how teachers’ perceptions vary by professional variables were found. Smith and Kovacs (2011) focused on the teacher variable, years of experience, and the student variables, socio-economic status (SES) and minority status. This study noted that teachers in high-poverty or high-minority schools and those with more years of teaching experience felt more strongly about the negative effects associated with NCLB policy.

Stuart Hammer (2004) examined accountability and standards-based reform in middle school science in the year after the state implemented high-stakes testing. Specifically, the study examined whether science teachers’ attitudes about accountability and standards-based reform varied by years of teaching at their current school, teachers’ attitudes toward school administration, percentage of students on free and reduced price lunch, school community environment (urban, suburban, or rural), percentage of minority students at a school, and grade level taught. The results indicated that “only the degree of support teachers receive from the administrators at their school had a consistent influence on the six factors of school reform examined” (p. 94). More support by administrators was associated with more positive teacher attitudes.

Limited doctoral dissertation work included Lowry (2010) who found teachers in high-stakes testing environments spent more time on test preparation than teachers in non-high-stakes environments and that this did not vary by gender or tenure. Mowers’ (2010) dissertation study examined if teachers under the consistent stress of accountability and NCLB showed signs of burnout and whether these signs varied by gender, teacher education, and years of teaching experience. This study noted that “female teachers exhibit slightly more emotional exhaustion and even less personal
accomplishment than male teachers and teachers with the most education exhibited the highest level of depersonalization toward their students” (p. 53). Anderson (2012) in his doctoral study examined how state policy influenced science teachers’ perceptions and how these varied by grade level taught, years of experience in science teaching, background in science and science teaching, and teachers’ perceptions of district leadership and policies. There generally was no variation other than “where teachers saw significant district influence they also reported more significant influence of state standards and testing on their teaching. They did not, however, see increased district influence as pushing them away from or toward inquiry or lab-based instruction” (p. 153).

**Conclusions and Implications**

This literature review identified science teachers’ perceptions of high-stakes testing and accountability, highlighted science teachers’ perceptions of high-stakes testing and accountability in relation to issues of student diversity and equity, and explained how science teachers’ perceptions of high-stakes testing and accountability varied by teacher demographic and professional variables. Key findings in the literature with regard to each of the three issues are highlighted. Then, implications for a research agenda are discussed.

Teachers have identified some positive effects of high-stakes testing, such as more rigorous curriculum and better alignment of curriculum with standards and assessments. However, teachers in general, and science teachers more specifically, negatively perceive high-stakes testing and accountability and its impact on both teacher practices and teacher well-being. Most significantly, teachers feel that the message of
what and how to teach demanded by high-stakes testing and accountability is in sharp contrast to the message of what and how to teach that is being delivered by state science standards and their personal beliefs. This tension may explain why teachers also perceive high-stakes testing and accountability to increase their level of stress and decrease their sense of autonomy. High-stakes testing alone appears not to improve or harm student learning. How teachers adjust to the demands and pressures of high-stakes testing, however, may have a significant impact on learning. This literature review suggests that teachers’ perceptions of high-stakes testing and accountability, rather than their views of standards, are a more significant determinant of their choice of practice.

One of the most significant objectives of high-stakes testing and accountability is the increased attention on achievement gaps and the needs of marginalized student groups. This literature review focused on the ELL population that currently makes up 11% of students in U.S. classrooms (NCELA, 2011) and is recognized as the fastest growing student population in U.S. education (U.S. Census Bureau, 2010). Large-scale tests show lower reliability for ELLs than for non-ELLs and test items with greater linguistic complexity tend to show greater bias against ELLs. High-stakes tests designed for monolinguals are not able to measure bilingual test takers’ knowledge (Penfield & Lee, 2010; Valdés, 1994). Teachers also perceive that high-stakes tests are inappropriate for ELLs. Bilingual education teachers perceive that high-stakes testing is impeding their effectiveness to teach and restricting the opportunities available to their students.

There was limited research found on whether there is variation in science teachers’ perceptions of high-stakes testing and accountability by teacher demographic and professional variables. This existing gap merits further study. By extension, it would
be valuable to see if any significant differences in teachers’ perceptions exist between teachers at high- and low-achieving schools where percentages of ELLs and other non-dominant groups differ and if they differ by teacher demographic and professional variables. Stuart Hammer (2004) found that teachers’ perceptions of administrative support were the most significant predictor of science teachers’ perceptions of standards-based reform and high-stakes testing. Further research may examine how teachers’ perceptions of high-stakes testing and accountability vary by teachers’ perceptions of their school administration and whether there is any variation by demographic or professional variables. As student demographics continue to change, it would be valuable to identify whether teachers’ perceptions of high-stakes testing and accountability vary by student demographic variables. Ultimately, examining the effect of teachers’ perceptions of high-stakes testing and accountability on teacher practice and student learning would be a significant contribution.
References for Appendix A


Rodgers, P. (2006). *What goal is of most worth? The effects of the implementation of the Texas Assessment of Knowledge and Skills on elementary science teaching* (Doctoral dissertation). Available from ProQuest Dissertations and Theses database. (UMI No. 3219180)


Appendix B

Teacher Questionnaire: High-Stakes Testing and Accountability Section

We are delighted to work with you on the P-SELL Project. We would like to know your beliefs and experiences in teaching science to your students. Thank you for your participation.

IV. Accountability Policy

9. Please indicate the extent to which you agree or disagree with the following statements about FCAT testing and school grades.

<table>
<thead>
<tr>
<th>Statement</th>
<th>Strongly disagree</th>
<th>Somewhat disagree</th>
<th>Somewhat agree</th>
<th>Strongly agree</th>
</tr>
</thead>
<tbody>
<tr>
<td>a. FCAT comes too soon in the year</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>b. The Florida A+ Plan provides a reasonable policy for grading schools</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>c. The rewards for high performing schools are well deserved</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>d. The sanctions for low performing schools are too harsh</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
</tbody>
</table>

Note: If you have not taught science for 5th grade since the implementation of FCAT Science, please skip to Question 12.
10. Please indicate the extent to which you agree or disagree with the following statements about **FCAT Science**.

<table>
<thead>
<tr>
<th>Statement</th>
<th>Strongly disagree</th>
<th>Somewhat disagree</th>
<th>Somewhat agree</th>
<th>Strongly agree</th>
</tr>
</thead>
<tbody>
<tr>
<td>a. Sunshine State Standards in science provide good guidelines</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>b. There are too many science topics to cover for FCAT Science</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>c. Topics for FCAT Science are too difficult for 5th grade students</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>d. Topics for FCAT Science are too difficult for my students this year</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>e. FCAT Science provides a fair measure of students’ mastery of science content</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>f. FCAT Science provides a fair measure of students’ basic skills</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>g. FCAT Science preparation dominates science instruction</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>h. Schools should be held accountable for FCAT Science scores</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>i. FCAT Science scores do not accurately reflect students’ science knowledge</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>j. FCAT Science scores do not accurately reflect ESOL students’ science knowledge</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>k. FCAT Science scores are influenced by reading abilities</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>l. FCAT Science scores are influenced by students’ economic status (i.e., poverty or wealth)</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
</tbody>
</table>
11. Please indicate the extent to which you believe that **FCAT Science testing** increases or decreases the following.

<table>
<thead>
<tr>
<th>Instruction</th>
<th>Greatly decreases</th>
<th>Decreases</th>
<th>Increases</th>
<th>Greatly increases</th>
</tr>
</thead>
<tbody>
<tr>
<td>a. Others telling me <em>what</em> to teach in science</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>b. Others telling me <em>how</em> to teach science</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>c. Time spent in preparation for FCAT Science</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>d. Drill and practice for FCAT Science</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td><strong>Students</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>e. Anxiety about learning science</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>f. Access to additional help in learning</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>g. Misbehavior during science instruction</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>h. Motivation to learn science</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td><strong>Teachers</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>i. Autonomy in teaching science</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>j. Creativity in teaching science</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>k. Stress in teaching science</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>l. Morale in teaching science</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>m. Turnover among science teachers</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
</tbody>
</table>
Appendix C

Teacher Background Form

Please provide the following information regarding your background.

I. Background

1. Gender:  
   1. Male  
   2. Female

2. Ethnicity:  
   1. White, Non-Hispanic  
   2. Hispanic  
   3. Haitian  
   4. Black, Non-Hispanic  
   5. Asian  
   6. Native American/American Indian  
   7. Multiracial  
   8. Other

3. Languages Spoken:  
   a) Please circle native language(s):  
      1. English  
      2. Spanish  
      3. French  
      4. Creole  
      5. Other (specify) ____________________________
b) Please circle other fluent language(s):

1. English
2. Spanish
3. French
4. Creole
5. Other (specify) ___________________________________

4. Years Teaching:
   a) In what year did you begin teaching (e.g., 2009)? ________
   b) In what year did you begin teaching at this school (e.g., 2009)? ________

5. Degrees
   a) What is the highest degree that you hold?
      1. B.A. or B.S.
      2. M.A. or M.S.
      3. Multiple M.A. or M.S.
      4. Specialist
      5. Ph.D. or Ed.D.
      6. Other (specify) ______________________
   b) Are you National Board certified?
      1. No
      2. Yes

III. ESOL Background

6. Please specify your training in English for Speakers of Other Languages (ESOL)-check all that apply.
   a. _____ Bachelor's degree in ESOL
   b. _____ Master's degree in ESOL
   c. _____ ESOL Endorsement through college coursework
   d. _____ ESOL Endorsement through school district (META)
   e. _____ Grandfathered in through teaching limited English proficient students
   f. _____ No preparation for ESOL
### IV. Science Background

7. Please indicate the number of courses that you have taken at the undergraduate and/or graduate level in each of the following areas:

<table>
<thead>
<tr>
<th>Area</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6 or more</th>
</tr>
</thead>
<tbody>
<tr>
<td>a. Methods of teaching science in elementary school</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>b. Methods of teaching science in secondary school</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>c. Physical science</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>d. Earth/space science (astronomy, geology)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>e. Life science (biology, ecology, environmental science, marine</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>science)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>f. Other (specify)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>